1990

Improving data quality in the Enlisted Master File

Sablan, Susan R.

Monterey, California. Naval Postgraduate School

(2)

NAVAL POSTGRADUATE SCHOOL

Monterey, California

THESIS

IMPROVING DATA QUALITY IN THE ENLISTED MASTER FILE

by

Susan R. Sablan

March 1990

Thesis Advisor: William J. Haga

REPORT DOCUMENTATION PAGE

1a Report Security Classification: Unclassified			1b Restrictive Markings
2a Security Classification Authority			3 Distribution Availability of Report: Approved for public release; distribution is unlimited.
2b Declassification Downgrading Schedule			
4 Performing Organization Report Number(s)			5 Monitoring Organization Report Number(s)
6a Name of Performing Organization: Naval Postgraduate School	6b Office Symbol (if applicable): 37		7a Name of Monitoring Organization: Naval Postgraduate School
6c Address (city, state, and ZIP code): Monterey, CA 93943-5000			7b Address (city, state, and ZIP code): Monterey, CA 93943-5000
8a Name of Funding Sponsoring Organization	8b Office Symbol (if applicable)		9 Procurement Instrument Identification Number
8c Address (city, state, and ZIP code)			10 Source of Funding Numbers: Program Element No / Project No / Task No / Work Unit Accession No

11 Title (include security classification): IMPROVING DATA QUALITY IN THE ENLISTED MASTER FILE

12 Personal Author(s): Susan R. Sablan

13a Type of Report	13b Time Covered	14 Date of Report (year, month, day)	15 Page Count
Master's Thesis	From To	March 1990	107

16 Supplementary Notation: The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

17 Cosati Codes			18 Subject Terms (continue on reverse if necessary and identify by block number)
Field	Group	Subgroup	data quality, data integrity, EMF

19 Abstract (continue on reverse if necessary and identify by block number)

In this paper, the importance of maintaining the quality of information in the Enlisted Master File will be established. The Enlisted Master File is generated by the Navy Enlisted System, one of many applications which process data for the Manpower, Personnel, and Training community. To clarify what technologies, policies, and procedures can contribute to improved data quality, a framework for classifying these initiatives is developed. The data quality control environment of the Navy Enlisted System is then evaluated with respect to that framework. Two deficiencies in the data quality control environment are identified. One is the lack of techniques to measure the quality of data in the Enlisted Master File, and the other is the lack of comprehensive plans for data quality control for the data base which will be a successor to the master file. A technique for assessing data quality is then tested, but its application to the Navy Enlisted System was not successful in this limited study. Technologies which could contribute to enhanced data quality in the environment of the future are discussed, and a plan for actively managing data quality is proposed. In closing, specific recommendations for improving the current data quality control environment in the Total Force Information Systems Management Department are presented.

20 Distribution Availability of Abstract: ☒ unclassified unlimited ☐ same as report ☐ DTIC users	21 Abstract Security Classification: Unclassified	
22a Name of Responsible Individual: William J. Haga	22b Telephone (include Area code): (408) 646-3094	22c Office Symbol: AS HG

DD FORM 1473,84 MAR 83 APR edition may be used until exhausted
All other editions are obsolete

Approved for public release; distribution is unlimited.

Improving Data Quality in the
Enlisted Master File

by

Susan R. Sablan
Lieutenant, United States Navy
B.S., United States Naval Academy, 1981

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN INFORMATION SYSTEMS

from the

NAVAL POSTGRADUATE SCHOOL
March 1990

Author: [redacted]
Susan R. Sablan

Approved by: [redacted]
William J. Haga, Thesis Advisor

[redacted]
James N. Eagle, Second Reader

[redacted]
David R. Whipple, Chairman,
Department of Administrative Science

ABSTRACT

In this paper, the importance of maintaining the quality of information in the Enlisted Master File will be established. The Enlisted Master File is generated by the Navy Enlisted System, one of many applications which process data for the Manpower, Personnel, and Training community. To clarify what technologies, policies, and procedures can contribute to improved data quality, a framework for classifying these initiatives is developed. The data quality control environment of the Navy Enlisted System is then evaluated with respect to that framework. Two deficiencies in the data quality control environment are identified. One is the lack of techniques to measure the quality of data in the Enlisted Master File, and the other is the lack of comprehensive plans for data quality control for the data base which will be a successor to the master file. A technique for assessing data quality is then tested, but its application to the Navy Enlisted System was not successful in this limited study. Technologies which could contribute to enhanced data quality in the environment of the future are discussed, and a plan for actively managing data quality is proposed. In closing, specific recommendations for improving the current data quality control environment in the Total Force Information Systems Management Department are presented.

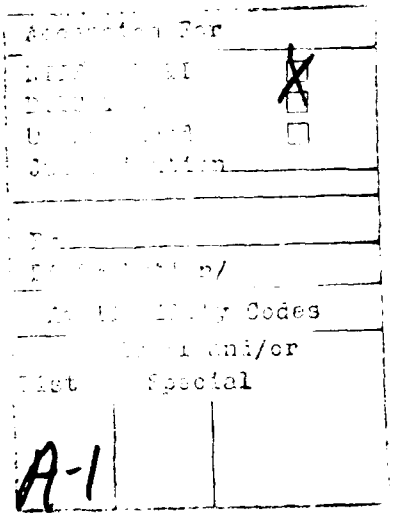


TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

The Navy must properly manage its personnel to successfully accomplish its objectives. In the 1990s, a time of shrinking fiscal resources, effective strategic planning for manpower, personnel, and training is absolutely essential. Reliable assessments of overall personnel strength must be provided to develop recruiting and accession plans. Accurate information regarding the skills of the current force is needed to set training quotas. Mobilization readiness and the ability to recall large numbers of personnel must be maintained. In the technological world in which the Navy operates, it must make effective use of each of its specialists. Tracking their qualifications is critical. Comprehensive information about language competency may be pivotal in a crisis situation. Incorrect security clearance information cannot be tolerated. Benefit and educational programs must be managed to a man. Whether measuring the personnel profile of the entire Navy for planning, or tracking the career of an individual sailor, accurate personnel data is an absolute necessity.

The Navy's managers of this manpower, personnel, and training information are well aware of the importance of their role in meeting broader service-wide strategies. The goals of the customer service center they operate are depicted in Figure 1. In support of their customers:

> All data/information initiatives are intended to meet the overall goal of providing timely and accurate information support to users and decision makers in order to increase the Navy's fleet readiness and mobilization ability. (CIRMP, 1989, p.3-3)

However, the information managers have not yet devised programs which provide a level of data quality which meets their customers' needs. This is due in large part to the age and complexity of the current systems.

A. THE QUALITY OF PERSONNEL DATA NEEDS IMPROVEMENT

The Navy currently collects its enlisted personnel data through an intricate network of interfacing systems. Some data are gathered interactively, with data validation performed as they are input. Other data are transmitted through the message system and consolidated by a central facility. The data are then processed by the Navy Enlisted System (NES)[1], a personnel accounting application developed in 1973. NES is run in a

[1] To aid in readability, Appendix A contains a list of acronyms with their long names.

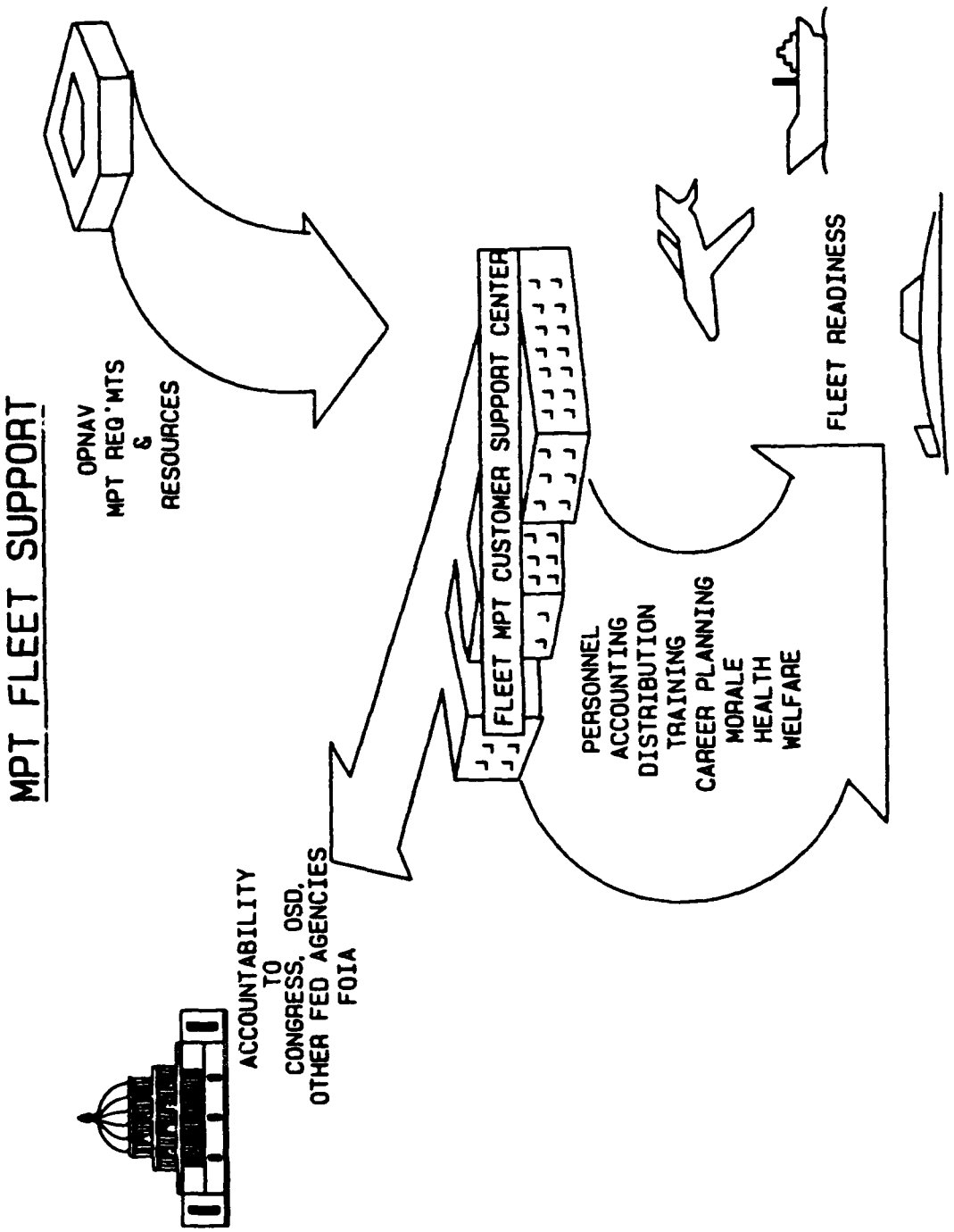


Figure 1. MPT Customer Support Responsibilities: This figure was taken from a briefing provided by OP-16.

batch environment, with updates five times a week. Various tapes, reports, and summaries are produced for use by managers of Manpower, Personnel, and Training (MPT) organizations. Individual sailors' records are used by assignment officers, Chief Petty Officer Selection boards, and a host of others. Though the quality of the data in the Enlisted Master File (EMF) which NES produces has improved over time, it is still an issue. In fact, during a 1989 conference of MPT functional users, the first concern identified was the "Accuracy and timeliness of data, [and] responses to requests & reports." (CIRMP, 1989, Figure 3-01, p.3-6)

This requirement for better data is not without precedent elsewhere. The government has passed legislation which requires that stored data be maintained:

> The Privacy Act of 1974, along with other federal regulations, has firmly established the importance of maintaining accurate, complete, and unambiguous information in computerized record systems. (Laudon, 1986, p.4)

In 1983, data integrity and quality was not ranked in a study of information professionals' top concerns. However, by 1986 it ranked 22nd, higher than such issues as decision support systems, computer graphics, and relational data base management systems (Brancheau, 1987, p.23). Business managers too, feel that stored data need to be accurate "Given a choice, managers have a strong preference for improvement in quality of information over an increase in quantity." (Davis, 1985, p.215) Those in the field of information science have begun research to measure the data quality problem. Mahmoud and Rice conducted a study of database vendors, and found the overall quality of the data they provided to be wanting in several areas, particularly in how vendors actually check accuracy and deal with outliers. They concluded that "Database accuracy is an important area of study because successful planning and business decisions depend on accurate forecasts, which, in turn, depend on accurate data." (1988, p.249) This conclusion mirrors that reached by the MPT functional users.

One important initiative which the MPT Information Resource (IR) managers feel will improve the quality of EMF data is the transition of the data structure from a flat master file to a data base. While moving to a data base technology will certainly facilitate data sharing and eliminate redundancy, it will not resolve all problems with data quality. Continued vigilance will be necessary. In the early 1980s, Brodie acknowledged that "Database reliability and integrity are poorly understood. In fact, data quality maintenance is a more severe problem than program reliability." (1980, p.246) This issue will continue to be of importance throughout the 1990s, as Martin described:

> We can draw a picture of the computing facilities of a typical future corporation. There will be one or more large computer centers--in many cases about the same number as there are today but with faster computers. These centers will be interconnected by telecommunications and will be jointly on line to most parts of the corporation. They will perform those computing operations which still benefit from centralization rather than distribution, for example, large number-crunching operations, large-scale print runs or printing needing special equipment, maintenance of files which are by their nature centralized, running old centralized applications which have not yet been converted to distributed form, and (particularly important) ***the maintenance of corporate data bases.*** (1981, p.23) (emphasis Martin's and mine)

Far from solving all data quality problems, data base technology may only exacerbate them:

> The need for accurate and complete data increases as more uses are made of those data. An accurate and complete data element in a dedicated system only affects that system, but in an environment where multiple users use the same data, the problem can be much more acute. The advantage of data base can only be achieved when the integrity of the data base can be ensured. (Perry, 1983, p.50)

Those responsible for MPT Information Resources Management (IRM) must keep in mind that "The rapid evolution of database concepts has been accompanied by the development of increasingly complex information systems with correspondingly complex data quality problems." (Brodie, 1980, p.253) Unfortunately, there is not a great deal of research about how to improve the quality of the contents of a master file or data base. Most of the research has focused on how to ensure that data is not corrupted during the collection, transmission, and storage processes. MPT IRM managers are well aware of these issues, and have adequate resources to resolve them. Technical expertise is available from contractors. While ensuring the integrity of these processes is an important part of a data quality control program, MPT IRM managers must also consider the timeliness, completeness, and accuracy of the actual data values.

This thesis will survey the current state of data quality control for the NES and provide recommendations for its improvement. This will be accomplished by building a framework for classifying data quality initiatives, by describing where various techniques fit into that framework, and by applying the framework to the NES environment. From that exercise, two major deficiencies in the data quality control environment of NES will be identified, and methods for eliminating these deficiencies will be evaluated. Lastly, recommendations for improving data quality control will be proposed.

To establish the relative importance of maintaining quality data, an issue easily ignored by managers in a world of competing priorities, this chapter depicts how the need for improved data quality has permeated the strategic planning of the IRM organization

responsible for the NES. Though the issue of data quality is addressed in numerous ways, there is no cohesive plan for ensuring that quality will improve. This is not to say that nothing has been done, many initiatives undertaken by MPT IR managers have improved data quality. What is missing is an overall plan that has as its primary goal, *improved EMF quality*. Lastly, this chapter describes why the management of the data quality control program must continue to be overseen by the IR managers.

In Chapter II, the framework for classifying data quality initiatives is developed. This framework can be used to assess the current state of data quality control in an organization, and will be applied specifically to the NES environment. To make the discussion of the framework more meaningful, basic terms are defined and the evolution of the framework is discussed.

In Chapter III, two deficiencies in NES data quality management are identified, by surveying the organizational controls present with respect to the framework established in Chapter II. These deficiencies are the inability to accurately assess the data quality of the EMF, and the need for enhanced data management when NES transitions to a new system. Resolving these deficiencies is by no means trivial. In order to establish some perspective on the scope of the challenge, the organizational environment and software characteristics of NES are described.

In Chapter IV, the problem of assessing data quality in the EMF is addressed. An approach which uses statistical decision theory is tested (Morey, 1982), and difficulties in implementing the technique are explored. An integer programming model which uses these assessments of data quality, along with a number of other parameters, to allocate resources to data maintenance techniques (Ballou and Kumar.Tayi, 1989), is briefly discussed.

In Chapter V, a plan for managing the quality of the data in the new personnel data base is proposed. This plan is a general one which could be applied to any new system which is data intensive.

In the last chapter of this thesis, specific recommendations for better managing the data quality of the EMF are summarized.

B. MPT IRM PLANNING EMPHASIZES IMPROVED DATA QUALITY

Managers of the MPT IRM function realize that improved data quality is an important strategic objective, both for the MPT business and the IRM organization. References to its crucial role are abundant in planning documents. The paragraphs below point out how essential improved data quality is, and how the MPT IR managers

have addressed that concern. While no specific plan for improving data quality has been developed, many initiatives list better data quality as a potential benefit. Since the MPT community is extensive, two lead organizations have been tasked with acting as the Chief of Naval Operation's (CNO) MPT IRM agent.

In order to clarify how MPT IRM strategies are conceived and implemented, the relationship of the lead organizations is described. The CNO's division for Total Force Information Resources and Systems Management (OP-16) sets policy regarding what IRM issues the MPT community will pursue. The CNO's organization is referred to as OPNAV, and its divisions are called OP codes. The Total Force Information Systems Management Department of the Naval Military Personnel Command (NMPC-16), who works for the Chief of Naval Personnel (CNP), must implement this policy. The OPNAV and NMPC houses of the organization work together very closely, and the head of these organizations is the same individual. For more information on the various roles of these organizations and others see the "Manpower, Personnel and Training (MPT) Information Resources Management (IRM) Program" (OPNAV Instruction 5230.22, 1986) or *The MPT Information Resources Management Strategy, Volume I: Executive Overview* (MPT IRM, Volume I, 1988, p.C-1).

OP-16/NMPC-16 take their guidance from the Navy's overall information management policy. This IRM policy is spelled out in the Secretary of the Navy's (SECNAV) instruction entitled "Department of the Navy (DON) Strategic Plan for Managing Information and Related Resources (IRSTRATPLAN)" (SECNAV Instruction 5230.10, 1987). In addition, the program guidance and reporting requirements are detailed in another SECNAV Instruction called "Information Resources (IR) Program Planning" (SECNAV Instruction 5230.9A, 1985). To support IR Program Planning certain activities must produce a Component Information Resources Management Plan (CIRMP). As the MPT IRM agent OP-16/NMPC-16 produce the CIRMP.

The CIRMP provides an overview of what the MPT IRM organization is doing to manage information as a strategic resource. Improved quality is a recurrent theme in these plans. One of 16 MPT IRM long range strategies is to "Use IRM principles throughout the MPT community to insure quality and valid data." (CIRMP, 1989, p.1-17) This is to support one of 29 major MPT business initiatives which is to "Improve Information Quality and Timeliness." (CIRMP, 1989, Figure 1-09, p. 1-16)

One of CNP's goals is to make better use of resources across the board. To that end, he has introduced Total Quality Management (TQM) to his organization. The

CIRMP shows that the MPT IRM community wants TQM principles to influence the quality of information products:

> Traditionally, product quality (whether manufactured products or information products) had been addressed reactively with programs like Quality Assurance. This type of strategy focuses on product improvement through inspection and error detection after completion. The TQM approach is "proactive:" it focuses on the achievement of product quality through the continuous improvement of all the component processes which in their totality, determine the quality of the product. (CIRMP, 1989, p.3-12)

The CIRMP is replete with these references to TQM and the philosophies it espouses. It also relates TQM initiatives to data quality initiatives. "Data standards, the Data Quality Assurance Program, and the Total Quality Management (TQM) process provide the framework for improving and maintaining the quality and integrity of MPT data." (CIRMP, 1989, p.2-4) However, throughout the document no plan for using specific TQM initiatives to improve data quality is set forth.

The data/information initiatives mentioned above, must be implemented in an environment filled with many constraints. In fact, "Manage[ing] IRM in a fiscally constrained environment" is listed as one of eight key IRM directions for CNP's organization. (CIRMP, 1989, p.3-1) Three of the 11 assumptions and constraints listed among CNP's chief directions and trends are related to the increasingly difficult funding environment. These are that "There will be increasing competition for less personnel and funding resources," that "The role of management will increase in order to attempt to effect savings in personnel and money," and that "There will be more and more 'micromanagement' from senior managers and organizations, including OSD [Office of the Secretary of Defense], the Office of Management and Budget, and Congress." (CIRMP, 1989, p.3-2)

What is the main point of this confusing conglomeration of issues, goals, and strategies? It is that data quality has consistently been a concern of managers, and that a comprehensive plan for data quality improvement must be developed. Currently, MPT IRM managers have undertaken many initiatives to improve data quality, but an overall data quality control program has not been developed. As the defense budget continues to shrink throughout the 1990s, military managers must be able to do more with less. If improved data quality will facilitate this, it must remain a top priority.

C. NMPC-16 MUST TAKE A LEAD ROLE IN DATA QUALITY

While information managers, functional users, and systems auditors all play a role in ensuring that quality data is maintained, the lead role must be taken by IR managers. Within the MPT IRM community, there is a strong trend towards expanding user responsibility for data quality. This trend is fine, as long as NMPC-16 realizes that they must continue to take the lead, by providing the users with the methods necessary to assess and improve the quality of the data for which they are responsible. Specifically, NMPC-16 cannot afford to assume that the data maintenance function will be shouldered entirely by users in the future. The paragraphs below explain current trends and detail why NMPC-16 must continue to plan and budget for a data maintenance activity.

In recent years, with the advent of on-line systems and end-user computing, much discussion has focused on data quality and the role of the user. The trend has been to shift some of the responsibility for maintaining data to users. This phenomenon has taken place because the user now has more access to the data base, is more familiar with it and what it means, and probably can make a better determination of its veracity. However, data maintenance will remain a shared concern:

> The integrity of the contents of the data base is the joint responsibility of the users and the data base administrator. The data base administrator is concerned more about the integrity of the structure and the physical records, while the users are concerned about the contents or values contained in the data base. (1983, p.91)

Perry further elaborates on this concept, explaining that the tasks listed below, are those which will help to maintain data quality. IR managers must:

1. Identify the method of ensuring the completeness of the physical records in the data base.
2. Determine the method of ensuring the completeness of the logical structure of the data base (i.e. schema).
3. Determine which users have responsibility for the integrity of which segments of the data base.
4. Develop methods to enable those users to perform their data integrity responsibilities.
5. Determine at what times the integrity of the data base will be verified, and assure there are adequate backup data between periods of proven data integrity. (Perry, 1983, p.91)

What is interesting about Perry's approach is that it doesn't abandon users to maintain data quality without the support of the information managers. To a certain degree, NMPC-16 has started to implement a user-oriented strategy for the EMF's data main-

tenance. The data elements and their characteristics have been analyzed and standardized for inclusion in a Data Dictionary/Directory System (DD/DS). Along with this, responsibility for each data element has been assigned to a functional manager. Some rejected transactions are sent back to the input source for correction. However, NMPC-16 has no specific plans to develop quality assessment and maintenance tools for use by functional managers. NMPC-16 should not abandon its control of corporate data quality. While the users can and should play an essential role, the MPT Chief Information Officer should ultimately be accountable for the quality of data in the master files or data bases.

Auditors have also played a role in the maintenance of data quality. They are responsible for devising methods to evaluate the accuracy of stored data, perhaps because information systems professionals did not build these into systems in the first place. There is an entire profession dedicated to auditing Electronic Data Processing systems, and their skills need to tapped by information systems managers. In addition, information systems managers should begin to assess the quality of the data in their own systems. This is nothing more or less than good management. Auditors are fine, but each system must have an established set of data quality control techniques to be applied to the system regularly, accomplishing the following objectives:

> Internal control accomplishes three major objectives. First, the "methodology" is designed to insure [ensure] that the accounting system provides accurate, complete, reliable and up-to-date information for making of management decisions. Second, it is intended to insure [ensure] compliance with policy directives, and legal requirements. And finally it protects the organization from carelessness, inefficiency and outright fraud. (Neumann, 1977, p.11)

At this point, no regular auditing is done on NES data. In the future, NMPC-16 managers would be wise to establish internal controls for data quality.

D. CHAPTER SUMMARY

In this chapter, three central themes were developed. The first theme established the importance of improving the quality of data in the EMF. The fact that the need for improved data quality ranks high among all concerns of managers and users of Automated Data Processing (ADP) systems was discussed. The discussion closed with an outline of how this thesis would attack the problem of improving the quality of EMF data. Simply stated, the strategy for achieving improved EMF quality is as follows: a framework for classifying data quality improvement techniques will be developed and applied to the NES environment, holes in the application of that framework will be

noted, methods of plugging those holes will be proposed, and recommendations will be summarized. The second theme of this chapter emphasized that organizational objectives acknowledged the importance of accurate personnel data as an MPT strategic resource. However, it pointed out that in spite of stated goals and objectives, no explicit program to provide overall coordination for controlling data quality has been developed. Many MPT IRM initiatives have contributed to data quality, and the survey approach used in this thesis will document those which impacted on NES. The last theme in this chapter explored the trend to place responsibility for data quality on end users or auditors. This discussion served to dissuade MPT IR managers from relying too heavily on data base technology and end users to completely resolve their corporate data maintenance problems.

II. DATA QUALITY IMPROVEMENT

This chapter develops a framework by which data quality improvement techniques can be categorized and managed. To enhance the discussion regarding how this framework was devised, definitions and background that explain what data quality is and how it is maintained will be provided. Since the field of data quality and its maintenance does not seem to be as well-documented or as clearly defined as that of software quality and its maintenance, parallels between the two will be drawn. This provides a basis for comparison, pointing out that data quality can be further defined, quantified, and improved. The framework for classifying Data Quality Initiatives will be used in the next chapter to survey data quality control with respect to the EMF.

A. WHAT IS DATA QUALITY?

Data quality is not a term which has a standard definition among information systems professionals. Sometimes the term data integrity is used in place of data quality, sometimes it is used to mean something different. Date says that "the term 'integrity' refers to the accuracy or correctness of data in the database." He further explains that "Many systems that claim to provide data integrity are actually using the term to mean *concurrency control* instead." (1986, p.444) (Date's emphasis) Weber says "It [data integrity] is a state implying data has certain attributes: completeness, soundness, purity, and veracity." (1982, p.8) Later, he classifies data quality control as one of six elements necessary to maintain data base integrity:

> To maintain the integrity of the database, the database administrator must undertake six control measures: (a) definition control, (b) existence control, (c) access control, (d) update control, (e) concurrency control, and (f) quality control. (Weber, 1982, p.170)

He implies that data quality is but an element of overall data integrity, and along with Date, says that concurrency control is a component data integrity, but not the whole picture. When Martin addresses data integrity, he discusses issues such as consistency, locks, conflict analysis, transaction loss, and deadlocks (1981, pp.287-306). He does not mention accuracy or completeness of information. Since none of the researcher's descriptions precisely capture the subject of this study, a working definition of data quality is proposed. When the term data quality is used here, it means the degree to which

stored data represent actual events as they took place, or the degree to which stored data record facts from a definitive source.

B. COMPONENTS OF DATA QUALITY

Various attributes of data quality exist, but which ones are commonly identified? To determine this, a much simplified version of the approach taken by McCall and associates with respect to software quality components is used. By compiling many researcher's definitions of software quality components and comparing and categorizing them, McCall and associates identified 11 characteristics of software quality. These included: maintainability, flexibility, testability, portability, reusability, interoperability, correctness, reliability, efficiency, integrity, and usability (1977, p.3-5). With respect to data quality, Laudon used the three components record completeness, record inaccuracy, and record ambiguity to measure the quality of records in Criminal-History Systems (1986, p.6). Ballou and Tayer's model for assessing the quality of data and processes in *information systems* addressed:

> Accuracy (the recorded value is in conformity with the actual value), timeliness (the recorded value is not out of date), completeness (all values for a certain variable are recorded), and consistency (the representation of the data is the same in all cases). (1985, p.153)

Weber said that completeness, soundness, purity, and veracity are components of data quality (1982, p.8). Date mentioned the elements accuracy and correctness (1986, p.444). A Chief of Naval Operation's publication, the *MPT IRM Data Quality Guideline* names the components accuracy, timeliness, and completeness (DCNO, 1988). In Table 1 the occurrence of these elements is summarized.

Components which appear most often include: completeness, timeliness, and accuracy. Completeness is the easiest component of data quality to determine. It is simply a measure of whether all data which should have been recorded, were recorded. There can be some subtleties involved when measuring completeness. If the field or attribute being measured is an optional one, it may be impossible to determine whether it was intentionally left blank. Accuracy is difficult to validate, by any means other than a manual record check. Inputs may be a valid entry or code, passing all edits and not creating an error, and still be inaccurate. Another attribute of data quality is timeliness, this is usually defined as the length of time it takes for an actual event or fact to be recorded. For example, if a sailor enlists today, how long does it take to record facts such as his name and date of birth? Later when he is advanced, how long does it take to re-

Table 1. DATA QUALITY COMPONENTS

COMPONENT	LAUDON	PAZER & BALLOU	DAVIS	WEBER	DATE	NAVY
ACCURACY	X	X	X		X	X
COMPLETENESS	X	X	X	X		X
PRECISENESS	X		X			
TIMELINESS		X				X
CONSISTENCY		X				
CORRECTNESS					X	
SOUNDNESS				X		
PURITY				X		
VERACITY				X		

cord that event on the file? However, just timeliness does not adequately describe this *attribute of quality*. Perhaps an additional measure, similar to what Ballou and Pazer call timeliness, could be volatility. This would be defined as the length of time the data is expected to be accurate. The value for volatility would vary from data element to data element, as some remain constant for the entire life of the record, and others change with different frequencies.

C. AN EARLY CLASSIFICATION SYSTEM

The early research in the area of data quality dealt primarily with managing errors. This research focused on the input process, and how better controls there could reduce the occurrence of errors. In 1969, Varley wrote a paper with the following objective: "The purpose of the paper is to provide procedures for detecting and correcting data input errors introduced by the human observer." (1969, p.1) In his paper, Varley made frequent mention of the Standard Navy Maintenance and Material Management Information System, on which he based his study. He was able to generalize many of his findings, eventually developing "a model--for evaluating the various detection and correction alternatives" taking into account "The necessary relationships between data worth, accuracy and cost." (Varley, 1969, p.1)

At the time of Varley's research, not much was published regarding the data input problem. Initiatives for improving the quality of the data focused on character recognition, automatic source data collection, and on-line computer systems (Varley, 1969, p.44). Interestingly enough two of these three initiatives, character recognition and on-line computer systems, have played a role in the development and enhancement of the NES.

In order to better understand the process of error detection, Varley identified seven independent locations where it could take place. These are:

- Data Generator
- Data Checker
- Keypunch Location
- Local Computing
- Central Computing
- Data Systems Analysts
- Information User

He explained how personnel at each of these locations in the data collection, processing, and use chain can often detect and sometimes correct errors. In many systems today, especially those which allow on-line input of data, the data generator, data checker, and keypunch location are one in the same. This is true of the Source Data System (SDS), which collects data for the NES.

Having defined the locations where errors could occur, Varley also categorized error detection and correction procedures based on the resources used. This classification system, detailed below, is useful only for classifying error detection and correction procedures typically undertaken during the maintenance phase of the Systems Development Life Cycle (SDLC):

- System-manual Procedure Class - This class refers to using manuals which provide detailed procedures for collecting data.
- Manual-visual Procedure Class - This class refers to verifying data using catalogs or reference materials.
- Manual-EAM [Electronic Auditing Machinery] Procedure Class - The process of manually verifying input with the help of admissibility checking falls in this class.
- Computer-aided Validation/Admissibility Procedures - This class refers primarily to admissibility edits and relational checking between data elements.
- Computer-aided Statistical Procedures - "This class of procedures uses either statistical inference or probability techniques for estimating the presence of errors."

- Computer-aided Table Look-up Procedures - This refers to verifying input values against a table.
- Computer-aided Master File and Cross-Reference Table Procedures - These procedures are those which use outside files for validation and possibly correction of data. (Varley, 1969, pp.143-144)

D. CLASSIFYING TECHNIQUES FOR DATA QUALITY

If it is possible to identify where errors are generated, and to target a particular class of techniques to correct those errors, then what is missing to achieve effective data quality control? According to Varley, it was the concept of error priority and how to establish that priority. He felt that there should be a means to identify the highest priority errors or the most valuable data. As Varley defines it, "The concept of error priority can be stated as the difference between the worth of a data element when it is accurate and the worth of the data element when it is in error." (1969, p.101) Error priority then, is predicated on a more basic concept, the concept of data value. "That is, what price per unit of accuracy are the users willing to pay for accurate data?" (Varley, 1969, p.157) This question is all important, for it provides a framework for measuring how much it is worth to an organization, for a particular data element to be maintained at a specific level of accuracy. Once this is established, it can be weighed against the cost of doing error detection and correction techniques such as those explained above. Optimally organizations should be defining the worth of their data, and then allocating resources to maintain the data within specified tolerances. What has probably kept organizations from measuring data worth, after all Varley wrote his paper 20 years ago, is the difficulty in doing so. The problem that exists for many systems, and certainly for those that are the object of this study, is that no one measure of data worth will do:

> In most cases the value of data to users changes from data element to data element as well as from user to user. This is more the rule than the exception. The system designer therefore, must decide what level of accuracy should prevail for each of the data elements. It is quite possible and reasonable to assume that some data elements are easier to bring to a given level of accuracy than other elements. This may require the system designer to perform the cost analysis at the data element level rather than the system level. (Varley, 1969, p.164)

This can be a tremendous job in large systems; the EMF currently carries several hundred data elements, and its successor data base is designed to carry over 500 (LDM Report, 1989, p.1). Economic evaluation of the worth of each of these data elements would be difficult at best.

Since Varley's research, other techniques for improving data quality have been introduced. Now there are other "locations" where something can be done to control the presence of erroneous data. These locations are outside the error generation and correction process; they are not people such as the data generator or systems such as the central computer. These locations can best be related to phases in the SDLC. In an article by Brodie, "The role of data quality is [was] placed in the lifecycle framework. Many new concepts, tools, and techniques from both programming languages and database management systems are [were] presented and related to data quality." (1980, p.245) The article concentrated on the physical characteristics of the system, those relating to hardware or software. Brodie discussed such issues as data types, structured programming, data abstraction, DD/DS, and data definition languages. He then related them to six stages of software development which he defined as:

1. analysis and definition of requirements,
2. logical design and its specification,
3. implementation and design,
4. implementation construction,
5. validation and verification, and
6. operation, maintenance, and evolution. (Brodie, 1980, p.248)

Brodie described where various tools and techniques can be used in his SDLC framework, except for the maintenance phase, which he says is surveyed adequately elsewhere.

Using Varley's research to understand where errors are created, and where and how they can be corrected, and Brodie's classification scheme involving the SDLC, a comprehensive way to categorize techniques for achieving data quality can be developed. Rather than focusing strictly on ADP related issues as Brodie did, this framework also looks at IRM issues as well. It advocates using a philosophy similar to that Varley proposed, where data value is a driver. Further, data quality control and maintenance issues are addressed early in a system's design. In this way, internal controls can be designed at the same time the system is crafted. Data values, relative or otherwise, are established early, and then priorities for data quality enhancement are set from the systems inception.

In Table 2 the phase of the traditional SDLC, an Object-Oriented methodology, and an Information Engineering development strategy are related to new Data Quality Technique Classifications. To aid comprehension, in the explanation that follows, column titles appear in the same type face as they do in the table. The **Development**

Strategy and its associated ***Phase*** are listed in the first column, with the SDLC Phases of Testing and Maintenance, covering all three strategies. For some, this in itself may represent a leap of logic, but Brodie's solution to that issue works here as well:

> Most popular database approaches have only two stages, e.g.[,] infological and datalogical, which do not permit an appropriate separation of concerns, nor do they facilitate the integration of database with software engineering technology. However, the development of database application is a large software development project. (1980, p.248)

In the second and third columns Varley's **Error Detection Locations** and **Error Detection and Correction Procedures** are included to show how they typically address only the maintenance phase of a software project. New **Data Quality Technique Classifications** are provided, and are related to the **Development Strategy *Phases***. In the last column, **Questions to Ask/Issues to Resolve**, those concerns which should be addressed at that stage of the life cycle are listed. These include some original questions and some posed by previous researchers.

E. METHODS OF IMPROVING DATA QUALITY

Using the framework just developed, the following paragraphs explore initiatives that can improve data quality.

1. Methods Involving Engineering in Quality

Efforts to manage data quality need not be restricted to the maintenance phase of the SDLC. There are various methods which can be used to ensure data quality, even before an ADP system has started to collect the data.

a. The Data Base

In a data base, "The data records are physically organized and stored so as to promote shareability, availability, evolvability, and integrity." (Davis, 1985, p.502) The ability to share data makes it more accessible to users, so it can be used more regularly. If the data is seen more often, users will help to assess and possibly improve its accuracy. Increased integrity also means improved quality.

b. Data Dictionary/Directory Systems

A DD/DS can help to improve data quality in several ways. When establishing a DD/DS, data should be standardized. To facilitate data standardization, naming conventions must be developed. These conventions help to avoid redundancy and more specifically:

> The dictionary helps to enforce agreement on the definition of each field and its bit structure. It helps to avoid having different fields in different places with the same

Table 2. DATA QUALITY INITIATIVES

DEVELOPMENT STRATEGY *PHASE*	ERROR DETECTION LOCATIONS	ERROR DETECTION AND CORRECTION PROCEDURES	DATA QUALITY TECHNIQUE CLASSIFICATION	QUESTIONS TO ASK ISSUES TO RESOLVE
Systems Development Life Cycle ***Requirements Analysis***	Not Applicable	Not Applicable	**Engineering in Quality**	Can the data we collect be validated? Can they be maintained? What is the value of these data to the organization? How accurate must they be? How timely?
Object-Oriented Development ***Identify Objects***				
Information Engineering ***Identify Data for Strategic Goal***				
Systems Development Life Cycle ***Design***	Not Applicable	Not Applicable	**Engineering in Quality**	How will the data be maintained? How is data value being quantified? What are the tolerances for accuracy? for timeliness? Place this information in the DD/DS.
Object-Oriented Development ***Create Schema***				
Information Engineering ***Determine what data to collect***				
Systems Development Life Cycle ***Coding***	Not Applicable	Not Applicable	**Software and Hardware Quality Controls**	Resolve technical problems such as management of redundancy, concurrency, and consistency controls.
Object-Oriented Development ***Implement Schema***				
Information Engineering ***Determine how to collect data***				
Testing	Not Applicable	Not Applicable	**Software and Hardware Quality Controls**	Test the technical issues mentioned above.
Maintenance	Data Generator Data Checker Keypunch Location Local Computing Central Computing Data Systems Analysts Information User	System-manual Procedure Manual-Visual Procedure Manual-EAM Procedure Computer-aided Validation Computer-aided Statistical Computer-aided Table Look-up Computer-aided Master File	**Methods Involving Data Capture**	Are data being captured in a timely manner? Can timeliness be measured? Are data accurate? Can accuracy be measured? Can erroneous data be detected? Can erroneous data be corrected? What are the correction priorities? What techniques are available to assess overall data quality? How are resources for data quality improvement being allocated?
			Methods Involving Error Detection	
			Methods Involving Error Correction	
			Methods for Assessing Data Quality	
			Resource Allocation Techniques	

name (homonyms) and the same field having different names in different places (synonyms). (Martin, 1981, p.388)

In addition, if established properly, the DD/DS can provide information which will make data maintenance easier once the system is deployed:

> The DD/DS contains valuable audit trail information about the data. For example, it could describe in detail where and how the data is used, and it identifies what program uses the data, where it appears in the program, what it is used for, what its relationships are to other programs, and whether any transformations were performed on the data.
>
> The DD/DS contains information about the users of the data, who they are, what they do with the data, how they use it, and so forth. It describes the physical devices that process data, and documents the software that use it, such as a DBMS [Data Base Management System]. In addition, the DD/DS also contains information about the kind of data that is used by the programs, the users, the physical device, the DBMS - these are all entities described in the DD/DS.
>
> All this information is important when tracing incorrect data entry or unauthorized access into the data processing environment. Evaluating this information can help identify the extent of an error; and it may be possible to identify the person responsible for perpetrating the error, or illegally accessing the data base. These output facilities can enhance the users ability to use the DD/DS as an audit trail aid. (Leong-Hong, 1982, p.55)

2. Methods Involving Data Capture

a. On-line Input

How does on-line entry improve data quality? First, recording events as they occur will allow the data to reflect the organization's actual status more quickly, increasing overall accuracy. Second, error checking and validation procedures can be executed while the data are being collected. Then errors can be corrected on the spot, by the individual who is entering the event and is most knowledgeable about it. On-line entry can be followed by immediate or batch processing, with different advantages and disadvantages; see Davis for a more complete discussion (1985, p.139).

b. Distributed Data Processing

Distributed Data Processing (DDP) shares many of the same advantages as on-line entry. Essentially, on-line entry and data validation are a low-scale form of DDP. For an application to be highly distributed, portions of the master database must be maintained at different locations. Martin provides more insight on the advantages of DDP:

> DDP permits *data entry* to be moved back to the user departments. There are several advantages to this. User departments can be made responsible for their own input data, for the accuracy and completeness of the data, and for the timeliness of the entry. Validation can be done by the machines *as the data are entered*; this is

desirable because errors can be corrected immediately, while the source documents are available. The laborious step of key verification following key punching can be avoided. (1981, p.23)

3. Methods Involving Error Detection

In addition to the techniques for error detection and correction described by Varley, audit packages can aid in error detection.

a. Audit Packages

Sometimes errors can be detected by using software, besides the applications software, to check the master file. The efficacy of these packages depends on many factors, principally the type of data collected by the application. Numeric data is generally easier to audit. Neumann surveys the features of seven audit packages in a U.S. Department of Commerce Publication (U.S. Department of Commerce, 1977).

4. Methods Involving Error Correction

Varley's framework covered error correction techniques fairly well. However, new technology has provided for error suspense files.

a. Rotating Error Files

According to Benoit the following benefits are gained from using a rotating error or suspense file:

- Error rejects are controlled to prevent loss
- Error rejects are corrected by authorized personnel
- Error rejects are corrected on time
- Corrections are subjected to the same edit, validation, and update process as the original entries
- Separation of duties is maintained
- The audit trail remains unbroken (1979, p.28)

5. Methods for Assessing Data Quality

Measuring data quality is an important area, where few useful techniques have been proposed. If it is true that today's managers value quality data over more data, and if it is true that many systems do not measure the quality of their data nor what they spend to maintain it, then what is the first step in changing this situation? A logical step would be to try and get a handle on the quality of the data that exist in a data file or data base at present. If there are no quality standards or measurement techniques for monitoring data, how can those responsible for data maintenance know where they stand. In the software industry assessment has recognized benefits:

> As data are collected and projects are measured, an organization improves its understanding of the software development process within its environment, and therefore the overall process is improved. (Valett, 1989, p.137)

To further the parallel with software measurement techniques, consider the following. Productivity is currently measured based on lines of code, and errors are currently measured based on cost per error. Both of these measures are imperfect, containing biases against fourth generation languages in the case of lines of code, and biases against quality products in the case of cost per error (Abdel-Hamid, 1989). However, researchers have made the decision that even an imperfect measure is better than none at all. This philosophy should be adopted in the study of data quality, because until some attempts to quantify the data quality problem are made, it can never be better understood.

There are various ways to measure errors in software. These include seeding models, where:

> A program is randomly seeded with a number of known 'calibration' errors. Then the program is tested (using test cases). The probability of finding j real errors of a total population J (an unknown) errors can be related to the probability of finding k seeded errors from all K errors embedded in the code. (Pressman, 1987, p.460)

The way this is expressed as an equation is that if:

K = Implanted errors

n = Errors detected in testing

k = Implanted errors detected and

J = Total errors in the program

then N can be found by

$$J = \frac{n \times K}{k}$$

and j can be found by

$$j = J - K$$

(Abdel-Hamid, 1989)

Mathematical techniques are valuable, because they provide simple ways to quantify something which otherwise is not easily measured. While some mathematical models for measuring data quality have been developed, these have not gained as wide an acceptance as software quality measurement techniques.

Haber and associates used regression analysis to determine if various factors played a role in the amount of reporting to a large-scale information system done by operational Navy units. The final goal was "to address the prior problem of identifying good vs poor reporters as a first step in approximation to identifying good vs poor data." (Haber, 1972, p.458)

Ballou and Pazer proposed:

> A model which under certain condition can trace the propagation and alteration of errors in data items within information systems. It also handles the impact of faulty processing on data items within information systems. It produces expressions for the magnitudes of errors in selected terminal outputs. (1985, p.151)

However, for this model to be used, all data must be numeric.

In his article, *Estimating and Improving the Quality of Information in a MIS* [Management Information System], Morey focused on measuring data quality of a MIS used for manpower planning in the U.S. Marine Corps (1982, p.37).

6. Resource Allocation Techniques

Another development worth mentioning, is that researchers are exploring how to better allocate resources for data quality enhancement. Optimization routines seek to provide the data maintainer with information about which data maintenance or error correction techniques should be applied to which data set (Ballou, 1989).

F. THE STATE OF THE ART - DATA MAINTENANCE

It appears that many of the organizations tasked with maintaining large stores of data do not use a rationale such as that prescribed in the framework. Data quality maintenance is a hit or miss proposition, where it is being done at all. In a study of the accuracy of data bases marketed by vendors, Mahmoud and Rice found that "Almost 20 percent of the database suppliers responding to the survey indicated that they do not check the accuracy of the data they receive." (1988, p.248) If this is true of organizations who market their data for profit, what does that say about organizations who collect their data for internal use?

Perhaps data maintenance suffers from the same lack of attention that software maintenance has over the past. Researchers seem to agree that:

> Software maintenance has until very recently been the neglected phase in the software engineering process. The literature on maintenance contains very few entries when compared to definition and development phases. Little research or production data have been gathered on the subject, and few technical approaches or 'methods' have been proposed. (Pressman, 1987, p.527)

Maintenance related functions are certainly less glamorous than those that are development related; programmers typically start out in maintenance and are promoted to development. In addition, organizations may not realize that "The principle of entropy also applies to stored data. The maintenance of data quality requires continuous inputs of resources." (Davis, 1985, p.611) Managers have only recently recognized the benefits to be gained from engineering software to reduce the need for certain types of maintenance. Perhaps the same can be done to eliminate some data maintenance. Once an error is identified, it costs more to correct than if the product or information were perfect in the first place. In software this is true, because correcting errors can result in the generation of more errors. It is also true with data because of a number of factors, including such things as communications overhead and additional processing time. It isn't hard to imagine that large benefits might be reaped, if data quality could be engineered and data maintenance procedures improved.

In the future, data maintenance issues should be addressed early in the SDLC. Too often in the past, these factors were not addressed until the system was operational and the data was already being collected. Varley recognized the importance of comparing data maintenance costs with the value delivered, and more recent research has focused on making these determinations up front:

> In assessing the establishment of databases, an important factor to be considered is the probability that the integrity of the data (accuracy, completeness, etc.) can be maintained. The assessment should also consider the risk from a degradation of data quality. (Davis, 1985, p.611-612)

G. CHAPTER SUMMARY

In this chapter a framework for classifying all data quality initiatives was developed. To facilitate a better understanding of the framework, a working definition of data quality was provided, and the components of data quality were identified. Then, an early framework for classifying data quality initiatives particular to the maintenance phase of the SDLC was presented, and the new classification system was proposed. Using this system, developments which contribute to enhanced data quality were described. The framework developed here will provide a reference point for the survey of data quality controls in the NES environment, presented in the next chapter.

III. NES DATA QUALITY CONTROL

In this chapter, two deficiencies in the data quality control environment of NES are identified. They are revealed by surveying the organizational controls which exist, with respect to the Data Quality Initiatives framework established in Chapter II. These deficiencies, the inability to accurately assess the data quality of the EMF, and the need for enhanced data management when the EMF transitions to a new data structure, will be further addressed in Chapters IV and V. Since the organization which supports the NES is extensive, and the survey of NES data quality initiatives would not be clear without some background information, the organizational environment and software characteristics of NES are first described. To gain an understanding of how hard it is to assess the overall effectiveness of the extended organization which supports NES, difficulties in measuring the costs of maintaining EMF data and in measuring the quality of the data are discussed. The information presented in this chapter will provide a basis for understanding the complexity of data quality problems in the MPT support systems and the difficulty involved in eradicating these problems.

A. CHARACTERISTICS OF THE ORGANIZATION

1. The Total Force Information Systems Management Department

The Total Force Information Systems Management Department (NMPC-16) is responsible for providing IRM support services to the Navy's MPT community. In addition to supporting each of the components of MPT, the department also has financial management, mobilization, and even limited pay system responsibilities. In 1988, the department underwent a major reorganization. The purposes of this reorganization were to move towards a future structure headed by a Chief Information Officer and to facilitate end-user computing. "'Data' oriented functions have been separated from 'technology' oriented functions." (CIRMP, 1989, p.1-10) The paragraphs below explain the new organization in relation to the data quality function, which is shared by a number of divisions.

The Customer Support Division (NMPC-163) is responsible for providing the MPT community's information needs. They are involved in numerous functions, from developing small systems for end-user computing to assisting with the preparation of reports and the management of contracts (Director, Reorganization, 1988). Their mission statement also requires them to assess customer satisfaction.

The Data Management Division (NMPC-164) is tasked with "Implement[ing] and execute[ing] programs to maintain and improve the sufficiency, accuracy, timeliness, integrity, reliability, accessibility, and security of MPT data." (Director, Reorganization, 1988) Most of the programs for data management are now handled or should eventually be absorbed by this division.

The Corporate Data Systems Division (NMPC-165) is responsible for the maintenance of the applications programs which create data bases such as the EMF. Their tasks which relate to data quality include developing programming edits and ensuring that program changes do not adversely affect the data base.

The Field Personnel Systems Division (NMPC-166) manages the systems that collect the data at the input source. These are systems such as the SDS, which provides local activities with the capability to record events (such as activity losses or gains and withheld promotions) in an automated manner, as they occur. This division must ensure that their systems provide for timely and accurate data input, another key element in data quality maintenance.

Finally, the Technology Support Division (NMPC-167) manages the hardware, including telecommunications networks, needed to collect and process data. They contribute to data quality by facilitating such things as timely transmission and processing of NES inputs. This responsibility is shared with the Consolidated Data Center (CDC) in Cleveland, Ohio, where NES file maintenance and reports are actually processed.

An NMPC-16 organization chart appears in Figure 2.

2. The Data Management Division

While many parts of the NMPC-16 organization play a role in ensuring that the EMF data is of sufficient quality, the responsibilities of assessing and ensuring data quality fall primarily on the Data Management Division (NMPC-164). Two branches of this division are tasked with maintaining data quality, each in a different way. The Corporate Data Maintenance Branch (NMPC-1641) contains the Enlisted Research Correction Section (NMPC-1641E). This section is responsible for correcting transaction errors from the NES updates and for updating erroneous personnel records in response to field inquiries. The Data Implementation Branch (NMPC-1642) contains the Data Quality Program Section (NMPC-1642C). This section "Establishes and implements procedures to measure, maintain, improve, and report data quality." (Director, Reorganization, 1988)

The Data Quality Program Section acquired a full-time manager in November of 1989, and it is still in the process of building its staff after the reorganization.

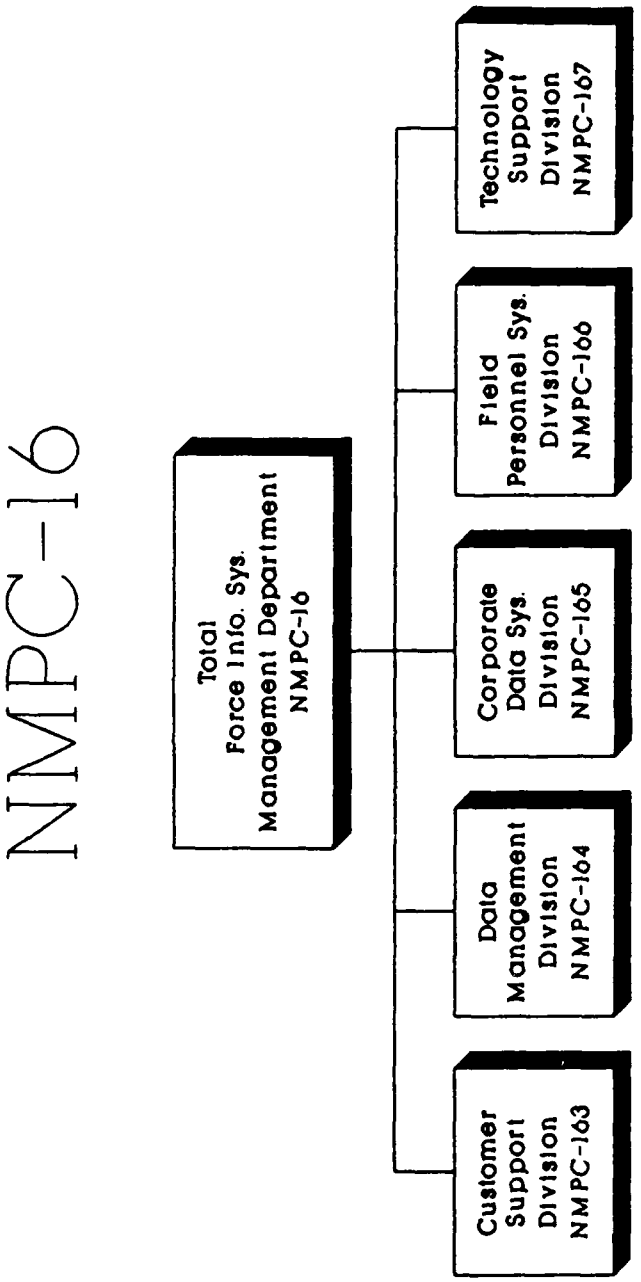


Figure 2. NMPC-16 Organization Chart

> Upon establishment this section will be responsible for Data Quality at the implementation level and will use the Data Quality Guideline as the basis for the preparation of implementation guidance for the echelon two and three commands under its purview. (CIRMP, 1989, p.4-9)

Managers would like this section to set goals and priorities according to need. Developing tools or techniques to measure data quality and to assign priority to data files or elements would facilitate this.

The Enlisted Research Correction Section's two basic functions deal with correcting specific instances of erroneous data. Their labor-intensive activities have been studied by the Naval Audit Service and by Troy Systems in recent years, in an attempt to make them more efficient. Some recommendations from these studies have been implemented and some have not. Later in this chapter, these recommendations will be summarized.

B. CHARACTERISTICS OF THE SOFTWARE

NES started functioning as the Navy's official information system on enlisted personnel in July of 1973. It is a transaction-oriented system, which is updated in a batch processing mode. Software developers used terminology and processes common to financial accounting systems, probably borrowing much of the design from those early systems. However, the development process was not simple:

> Systems functions were difficult to obtain as functional managers of the day had a poor understanding of how to define what the system should do. To a large extent system developers had to decide what the data requirements for the data base would be and what mechanisms would be used to collect the data. Furthermore, rules for editing and updating were also constructed by the ADP developers. (Milestone IV System Decision Paper, p.2)

Obviously, this has had serious consequences throughout the life of the system, and even today affects the quality of the stored data.

NES updates occur daily; special monthly, quarterly, and year-end processing is also done. Various inputs for the updates are collected in Washington DC. Then, they are bulk data transferred and processed further. NMPC-16 is billed for this processing which takes place at the CDC. The CDC supports much of the pay and personnel community:

> Hardware consists of IBM 3081 and 3084 mainframes, IBM 3380 and 3350 disk, and a 3851 mass storage system. This equipment is connected through high-speed data communications lines to remote centers at Washington, D.C. and Cleveland. The complex supports batch and interactive processing using IBM's MVS/XA operating system and Systems Network Architecture. (CNO, 1988, Appendix 9)

Specific software characteristics include the following:

- 260 modules
- 900 programs
- 400,000 lines of code
- Interfaces with 25 other systems
- Average of 50,000 transactions per daily update
- Average of 90 requests for changes or ad hoc reports are active at any time
- Edits are verification, rather than exception oriented
- It uses State and Country Code Tables, Navy Enlisted Classification Code (NEC) Tables, Rate Tables, Unit Identification Code Tables, Language Code Tables, Professional Pay Tables, and Submarine Pay Tables to validate data (Monroe, Slide Show)

C. CHARACTERISTICS OF THE DATA AND DATA BASE

The EMF which NES creates has the following characteristics:

- Averages 630,000 records
- Record Information:
  - A maximum of 3000 bytes
  - Average record length is 990 bytes
- Record data categories include:
  - Personnel Data
  - Rate/Rating
  - Skill Data (NEC's)
  - Service Data
  - Evaluations
  - Duty Preference
  - Current Activity/Duty Assignment
  - Availability
  - Assignment Data
  - Career History
  - School History
- Users of the data include:
  - Navy
  - Department of Defense

- Chief of Naval Operations
- Civilian Agencies (Monroe, Slide Show)

D. NES EXPENSES

1. Cost of Running NES

It is important that a quality data base be maintained, so that the resources used in collecting and processing the data are well-spent. The information in Table 3 should provide some perspective on the cost of NES. These values reflect only central processor, tape processing, and direct access storage device use. They do not include telecommunications, data collection, or other costs. They were provided by the Navy Finance Center (NFC) in Cleveland, and are an output of the Resource Accounting System (RAS). If the costs displayed for each month are added and averaged, an approximate monthly cost can be obtained. If that cost is multiplied by 12, the approximate yearly charge of NES File Maintenance and Reports together, is over a million dollars.

Table 3. NES COSTS FOR SEVEN MONTHS (1989)

Month	NES File Maintenance	NES Reports
March	$76,712	$31,545
April	$89,072	$40,328
May	$71,859	$31,563
June	$71,728	$33,072
July	$70,706	$39,980
August	$81,472	$32,618
September	$80,057	$51,676
Yearly(Avg.)	**$928,467**	**$447,055**

2. Costs Associated with Data Maintenance

Costs associated with maintaining data quality include researchers' salaries, processing costs, micro-fiche or document reproduction costs, and communications overhead (telephone calls and meetings). As a sample of these costs, the approximate salaries of the Enlisted Research Correction Section are displayed in Table 4. The civilian salaries are taken from a January 1989 General Schedule (GS) pay chart, assuming step five for each grade. The military salaries are taken from the basic pay chart effective

January 1989 (assuming E-8, 16 years; E-6, 8 years; E-5, 6 years; E-4, 4 years; E-3, 2 years).

Table 4. ENLISTED ERROR RESEARCH SALARIES

Old Organization		NMPC-1641E	
Grade/Rating	**Salary**	**Grade/Rating**	**Salary**
GS-204-9	$27,026	GS-204-9	$27,026
GS-204-7 (2)	$44,186	GS-204-7	$22,093
GS-204-6 (2)	$39,764	GS-204-6	$19,882
GS-204-5 (11)	$196,218	GS-204-5 (2)	$35,676
YN PNCS	$23,448	YN PNCS	$23,448
YN PN1 (3)	$48,348	YN PN1 (2)	$32,232
YN PN2 (3)	$42,336	YN PN2 (2)	$28,224
YN PN3 (2)	$24,984	YN PN3 (1)	$12,492
YN PNSN (3)	$30,888	YN PNSN (2)	$20,592
Total	**$477,198**	**Total**	**$221,665**

When attempting to identify costs involving data quality maintenance for NES, each cost had to be gleaned from a different source. Costs of processing came from NFC, manning information and transaction data from NMPC, and pay information from Civilian Personnel and Military Disbursing Offices. All this information was necessary, just to figure out a rough approximation of the cost of researching and reapplying one erroneous transaction. This cost is displayed in Figure 3. Of course, this figure does not include the overhead costs mentioned before. If an organization wants to get a better idea of whether or not it is beneficial to perform a particular data maintenance activity, they need to have a quicker way to make that assessment. The procedure for figuring out the cost of correcting a transaction required several months worth of effort to complete, due to the lack of readily available management information. Appendix B contains a summary of the statistics used to determine the averages provided in the figure.

E. USING THE DATA QUALITY INITIATIVES FRAMEWORK

Various measures are being taken in the MPT IRM organization to improve data quality. These efforts span a continuum from attempting to engineer in quality to developing more efficient ways to correct errors. Using the framework described in

YEARLY AVERAGES:	
NES TOTAL INPUT TRANSACTIONS	7,376,778
NMPC-16 ERROR RESEARCH TRANSACTIONS	238,781
NMPC-16 INPUT FILE CORRECTIONS	256,669
NES FILE MAINTENANCE PROCESSING COST	$928,467
ERROR RESEARCHERS SALARY COSTS	$221,665
COST OF PROCESSING ONE TRANSACTION	.13c
COST OF HANDLING ONE MAINTENANCE ACTION	.45c
COST PER TRANSACTION	.58c

Figure 3. Cost of Correcting one Transaction

Chapter II, the paragraphs below survey past and current initiatives targeted at improving data quality. From this, two deficiencies are identified. These are the inability to accurately assess the quality of the data in the EMF, and the need for enhanced data quality management when NES transitions to a new system.

1. Efforts Aimed at Engineering in Quality

Often the quality of data can be improved by better planning and design. As in any other quality assurance situation, it is usually cheaper to do something right the first time.

a. Information Benefit Analysis

Information Benefit Analysis (IBA) plays a role in improved data quality. The purpose of IBA is to identify solutions to business problems or to explore strategic opportunities, by studying the organization and how it functions. Once this has been accomplished, a data analysis is done. Data analysis is a nine step process which seeks to identify problems or opportunities, find solutions, and assess the value of these sol-

utions. Benefits identified may be tangible or intangible. (DCNO, IBA Guideline, 1989, p.1-12)

The CRIMP states that the "Information Benefit Analysis Methodology (IBA) is used to evaluate information needs and set priorities for allocation of resources." (CIRMP, 1989, p.2-27) The IBA is currently used before the first milestone of Life Cycle Management, that is before the requirements of a system are even determined. Since IBA forces the system planner to quantify information benefits (though not necessarily monetarily), some of these assessments can be carried over to the requirements analysis phase. Perhaps, using IBA, the relative value of data elements could then be determined.

b. Data Issue Resolution

Data Issue Resolution provides for consistency when data elements are to be added, changed, or deleted from a corporate data base. In NMPC-16 and throughout the MPT community, managers saw the need to have someone to oversee these issues:

> In addressing data integrity and continuity of operations, a centralized Information Systems Change Comptroller is required to ensure that data issue resolution which results in change is executed correctly, accurately, and in a timely manner. (CIRMP, 1989, p.4-15)

c. Data Standardization

Data element standardization has many benefits. Its primary focus is to facilitate better sharing of resources and to provide for more efficient support of the MPT community (DCNO, Data Element Standard, 1989, p.1). The *MPT IRM Program Data Element Standard* addresses such issues as Data Element (DE) Design, DE Definition, DE Naming, DE Approval and DE Registration. A corollary benefit to data element standardization is improved data quality. This is true for several reasons. First, interfacing with other systems does not require processing overhead to convert from one format to another, where potential mistakes or misunderstandings can occur. Second, the definitive source and the valid values for a particular data item are established, and then they are documented to prevent confusion and misinterpretation.

d. The Information Resources Encyclopedia

Information developed as a result of data element standardization is entered in the Information Resources Encyclopedia (IRE). A sample IRE entry is found in Figure 4. While it is sometimes referred to as an active system, the IRE currently is only a data base of standardized data elements. In the future, NMPC-16 plans to investigate

the possibility of an independent, active system. There are plans to enhance the IRE in the following way:

> A PC-based [Personal Computer], user-friendly front-end to the MPT and PAY IRE will be developed and implemented in 1990, and training in the use of the front-end will be provided to the MPT and PAY community [communities].
>
> The IRE will be populated with the IMPDB [Integrated Military Personnel Data Base - successor to the EMF] logical and physical design specifications; standardized, corporate IMPDB data elements; and MPT corporate data architectures in 1990. . . . The IRE will have an automatic update capability, with no manual data entry required. This will ensure that the corporate metadata repository, which documents the MPT corporate information system environment, remains valid and up-to-date. . . .
>
> A fully developed, two-way interface between the IRE and the corporate DBMS directors will be in place in 1993. The interface will automate the process of creating logical database designs and will generate a physical corporate database design from the logical design information. . . . (CIRMP, 1989, p.4-11)

If this automated development of data base designs includes validation routines, it would enhance the role the IRE plays in ensuring data quality. Then, all applications would generate their validation routines through the IRE. This would eliminate a significant problem, which is keeping the edits on all satellite systems in synchronization with those of the master system. It appears that NMPC-16 does have plans to standardize validation routines in the future, but it may not be done under the umbrella of the IRE (*though DD/DSs often have validation meta data*). In another document, which describes a three-level architecture for NMPC's future systems, the IRE is described as a corporate data base:

> This database will contain information necessary for the planning, design, development and maintenance of information systems. This information may be distributed across multiple physical databases and embedded in ADP management support systems. (Software Solutions, 1988, p.4)

Later, the document describes another layer of the architecture, System Utility applications, which are "those used to update and insure [ensure] the integrity of the corporate data." There are four of these systems mentioned, including a Table Maintenance System, an I/O (Input/Output) Validation System, a Transaction Processing System, and a Secondary Database Creation System (Software Solutions, 1988, p.4). If this strategy is used, three of the four systems named above will play a role in ensuring data quality. The Table Maintenance System will contain any tables required for validating input, the I/O Validation System will contain the admissibility edits, and the Transaction Processing System will contain the relational edits (those which must check data in the master data base to be performed) (Software Solutions, 1988, p.4). Whether the synchroniza-

```
                PRINT OF PAY-GRADE-V ENCODED IN N1653
00100           ITEM
00200           HELD-AS ALPHAN 2
00300           HELD 2 ALPHAN 1
00400           HELD 3 ALPHAN 4
02000           DESCRIPTION
02020             "THE PAY GRADE IN WHICH THE MEMBER IS CURRENTLY",
02030             "SERVING."
02100           CATALOGUE
02110             'ADMITS', 'NAMPS', 'MANCLASS', 'ACMP', 'JOB', 'MMRS',
02115             'NES-REPORT-430', 'NES-DAILY', '90-SERIES', 'NES-MONTHLY',
02120             'NRMIN', 'NITRAS', 'NIS', 'STP', 'JUMPS', 'MMRDB',
02125             'RT3S', 'RT3S-SUR', 'RT3S-AIR', 'RT3S-TM', 'NES-REPORT',
02140             'RESULTS', 'PEOPLE', 'MONEY', 'MMBAS', 'NES', 'NES-DAILY',
02150             'NES-REPORT-470', 'NES-REPORT-481', 'NES-REPORT-432',
02154             'NES-QUARTERLY',
02158             '69-SERIES', '86-SERIES',
02164             'NES-REPORT-428',
02167           '65-SERIES', 'NES-REPORT-427'
03100           COMMENT
03102             'CODE DESCRIPTION',
03104             '1    ENLISTED',
03106             '2    OFFICER',
03108             '3    WARRANT',
03110             '4    GS',
03112             '5    WAGE GRADE',
03114             '6    WL (LEADER)',
03116             '7    WS (SUPERVISOR)',
03118             '8    WD',
03120             '9    WN',
03122             '0    WA'
03146           "PERMISSIBLE-VALUES"
03148             'PAYGRADE   RANK   PAYGRADE RANK  '
03150             '   E1    SR    1  ENS   '
03152             '   E2    SA    2  LTJG  '
03154             '   E3    SN    3  LT    '
03156             '   E4    PO3   4  LCDR  '
03158             '   E5    PO2   5  CDR   '
03160             '   E6    PO1   6  CAPT  '
03162             '   E7    CPO   7  RADM   '
```

```
1989 12 04  09.03.10                    MANAGER SOFTWARE PRODUCTS                    PAGE    60
UPDATE STATUS N1653                          DICTIONARY SBPROJ

PRINT OF PAY-GRADE-V ENCODED IN N1653 (CONT)

03164             '   E8    SCPO  8  RADM   '
03166             '   E9    MCPO  9  VADM   '
03168             ' W1 WO    '
03170             ' W2 CWO   '
03172             ' W3 CWO   '
03174             ' W4 CWO   '
                END OF PRINT
```

Figure 4. IRE Entry for the Data Element Pay-Grade

tion of edits is achieved by the IRE or by the other system described, common edits, which can be simultaneously updated, should significantly improve data quality.

2. Efforts Aimed at Data Capture

Conventional wisdom about data quality indicates that it is most beneficial to validate information as it is collected. Editing or validating at the input source eliminates overhead and allows the most knowledgeable person to correct the data. Of course, this was not always possible in the early days of information processing.

a. The Source Data System

In the last decade, the SDS was designed and implemented at many shore Personnel Support Detachments in the continental United States. This system contains

a portion of the EMF (and other corporate data bases) called a mini-master. It allows offices responsible for personnel and pay accounting to update the master data bases in a timely and efficient way. Many of the edits or validation checks performed by the master file, are also performed at the local activity, allowing erroneous information to be corrected before it is recorded. Since there is some potential for differences between the master and the mini-master, and because not all edits can be done in local software, it is still possible for transactions to fail to apply to the EMF in a NES update.

b. Elimination of Optical Character Recognition

At one time Optical Character Recognition (OCR) was a new, promising technology. Apparently, its usefulness in Navy personnel accounting has been disproved. In many cases, the scanning process created additional errors. In fact, overall "the success of OCR has been limited. OCR devices are still relatively expensive and their reliability lower than other input devices." (Weber, 1982, p.221) Consequently, NMPC-16 has done away with scanning as much as possible.

c. Improving Timeliness

MPT IR managers have gotten together and established a task force to determine what represents an appropriate measure of timeliness. In addition, the task force is seeking to set standards for improving timeliness:

> We are currently revising the way timeliness is measured. Our goal for 100% [in a later document (Teter, 1989) this was changed to 99.5%] submission is 15 days. The goal includes a 5 day window for field preparation, 10 days for mailing and/or transmission of data from various input systems (DMRS [Diary Message Reporting System], EPMAC [Enlisted Personnel Management Center - responsible for DMRS input] SDS, and OCR). The timeliness statistics will be measured in calendar days from the date of occurrence. Future date transactions, retroactive starts, changes, UA(s) [unauthorized absences], wavier requests, NRA(s) [Navy Recruit Accessions] and corrections will be excluded. (Commander, Draft of Letter Subject: Timeliness Performance Report)

3. Efforts Aimed at Error Detection

There are several locations and methods used to detect errors. Those used in NES now, or planned for future use are described in the sections that follow.

a. Edits and Reconciliation

Errors are detected in several different ways. Transaction errors are caught by edits at the field in SDS, at EPMAC in DMRS, and during NES updates for all input systems. In addition, errors are detected through the use of reconciliations with other data bases such as the Social Security Administration's System or the Joint Uniform

Military Pay System (JUMPS). These reconciliations are conducted on a monthly basis. The JUMPS reconciliation:

> A. Provides assurance that data on the master record are identical to related data in the personnel system for the same member.
> B. Identifies the aspects of the update processing which may require modification to keep the financial system "in line" with the personnel system.
> C. Provides a periodic review of the validity of data maintained in the personnel system and forwarded to NAVFINCEN [NFC]. (JDRM, 1990, p.2-2)

In any situation where there is a no-match condition, a report is printed, and error research is conducted. In some cases, correction transactions are generated, and are processed in the next JUMPS update. Periodic queries and file sweeps are also done on an ad hoc basis to identify trouble spots.

b. NES Update Statistics

During each daily update, NES File Maintenance (F/M) Update Statistics are produced. These statistics tell how many transactions of each type were entered in the update, how many failed to process, and the error code each failed transaction was assigned. The reports also contain information on the age of certain records which are in an exception status. Those familiar with the statistics can often identify major errors or problems with an update. These problems would include things such as improperly sequenced or garbled tapes being processed.

c. On-line Error Trends

There has been discussion regarding the development of a data base of NES F/M Update Statistics. Then, analysis of this data base could be conducted to determine whether a particular update produced error rates within set tolerances. This would reduce the organization's reliance on "experts" who have been reviewing statistics for years, and would allow large processing errors to be caught more reliably.

4. Efforts Aimed at Error Correction

The last way to achieve data quality is to actually correct records identified as erroneous. This is the final check point for ensuring the accuracy of data. The paragraphs that follow provide a summary of NMPC-16's ongoing initiatives for improving the error correction process.

a. The NES On-line Correction System

NES now uses an automated suspense file to control rejected input transactions. As explained in the last chapter, there are a number of benefits associated with a rotating error file. The NES On-line Correction System (NOCS) is an interactive system which provides for on-line viewing of erroneous transactions. It also allows a

transaction to be generated and submitted into the daily NES update. The system is used by the Enlisted Research Correction Section to turn around transaction errors from the daily updates or to submit correction transactions.

b. Management Reviews

To improve the effectiveness of the Enlisted Research Correction Section, NMPC-16 solicited the help of two outside organizations. Troy Systems did several data quality reviews in late 1988 and throughout 1989, and in 1987, the Naval Audit Service provided an assist visit. The outcome of these investigations is briefly discussed.

(1) Troy Systems - September 1989. On September 29, 1989, Troy Systems, Incorporated completed a Data Quality Improvement Report for NMPC-1642. The work performed under this contract centered on resolving problems with two particular data elements, citizenship and place of birth. The report contained recommendations concerning restructuring edits and standardizing tables for coding inputs. One of the report findings indicated that the only place the data were output was to the Central Adjudication Facility, responsible for approving security clearances. NMPC-16 has a Memorandum of Understanding tasking them to provide support to this organization. (Troy, September, 1989, pp.1-10) The work done under the contract was obviously worthwhile, but the dollars spent on resolving that data quality problem might have been better spent on a need more pressing to the Navy's MPT community. When compared with all MPT priorities, improved security data might not deliver the most value to the MPT managers. Unfortunately, until data elements are assigned a value and data maintenance priorities are compared to one another, optimal allocation of data maintenance resources cannot be achieved.

(2) Troy Systems - July 1989. On July 7, 1989, Troy Systems, Incorporated completed a report on error research and correction analysis, for NMPC-164. It contained a number of useful recommendations which are detailed in the paragraphs that follow.

Their first recommendation was that a section be established within NMPC-1641 to do data error analysis, instead of just error correction (Troy, July 1989, no page numbers). At present, the Data Quality Program Section is just getting staffed up. The functions that would be done by a data error analysis section could also fit here, in NMPC-1642C, depending on at what level of detail NMPC-16 managers want to split the data maintenance function.

The next recommendation was that the NOCS be enhanced (Troy, July 1989). While the name NES On-line Correction System seems to indicate that the

corrections are made on-line, they are actually batched and processed in a regular daily NES update. The nature of the recommendations for NOCS enhancements varied. Some were designed to give the manager of the error research section better productivity statistics with which to measure researchers. Others were crafted to provide transaction statistics by input source, and still others were geared to improving researcher productivity by providing more tools.

The report also recommended that the Enlisted Research Correction Section "standardize procedures documentation" (Troy, July 1989). While the report did not elaborate any further on this subject, it appears to be a valid recommendation. All of the researchers who correct erroneous transactions are civilians, who would have little knowledge of Navy or ADP terminology when hired.

A function which consumes a great deal of time in NMPC-164IE is providing error correction assistance to field activities and individuals who call for help. The Troy report recommends that some of these functions be transferred to NMPC-163, the Customer Support Division (Troy, July 1989).

Currently, error research for correcting transactions must be done by using either paper or microfiche transaction reports. A recommendation to provide an on-line transaction file was included in the Troy report. The officer system already has this capability, and the enlisted system could benefit from it as well. Questions such as how many months/years of transactions can be stored will need to be resolved, as the volume of the enlisted system far exceeds that of the officer system.

The report included a recommendation to move transaction error correction to the source where the transaction was created (Troy, July 1989). This is a useful recommendation in many cases, but only with a certain note of caution which was not mentioned in the report. Well-meaning managers might see this as a panacea, as a way to eliminate the transaction correction function all together. In fact, cuts to manning have already been based partially on the assumption that in the future, the need for transaction error correction at headquarters would be greatly reduced. However, as long as there are edits, and there is potential for the edits to differ between system, transactions will error out of the master file update. Not all of these transactions can be more easily corrected at the input source.

The report provided two final recommendations. These were that the personnel system to pay system discrepancies be analyzed further and that all OCR transaction processing be eliminated. The department is actively implementing these recommendations.

(3) Naval Audit Service - 1987. In May of 1987 the Naval Audit Service completed a Management Consulting Report for NMPC-16. At that time, before the reorganization, "NMPC-1654 was the branch responsible for maintaining the officer and enlisted master file (OMF and EMF) for the Manpower [Personnel] and Training Information Systems (MAPTIS)." (Hickman, 1987, p.1) The executive summary of this report provided the following analysis:

> NMPC-1654 exerts an enormous manual effort to maintain data in the automated information systems. While data quality assurance certainly requires some human oversight, there are several areas within NMPC-1654 in desperate need of automation. There are also efficiencies to be gained by reorganizing the sections, changing work procedures, and improving the work environment. (Hickman, 1987, p.1)

An issue addressed twice in this chapter already, who is looking out for error trends and how are they doing it, appears in this report as well. In 1987, there was a section dedicated to this function. It was the Data Systems Analysis Section (NMPC-1654C), which contained one Lieutenant, one other military, and one GS-12 civilian. "This section researches[ed] and analyzes[ed] the data in the master files, looking for error trends and anomolies which could erode the validity of the data base." (Hickman, 1987, p.7) However, Hickman questioned their method of setting priorities:

> NMPC-1654C plays an important role in actively searching for problems in the database; however, the guidance on what to research is mostly self-generated. There are no formal guidelines or priorities for error detection, and therefore, the section appears to be in a reactive mode when problem solving. (Hickman, 1987, p.10)

Hickman also says when error trends are identified and programming changes are necessary (as frequently they are), it would be more efficient to implement corrections if quality assurance and applications programming personnel reported to the same boss (Hickman, 1987, p.11). This was not the case in the department then, and is not the case in the new organization. The suggestion has merit, for reasons beyond those in the report. Often, isolation of problems requires programming intervention, this is likely another reason why some data maintenance functions are still being done by the Corporate Data Systems Division today.

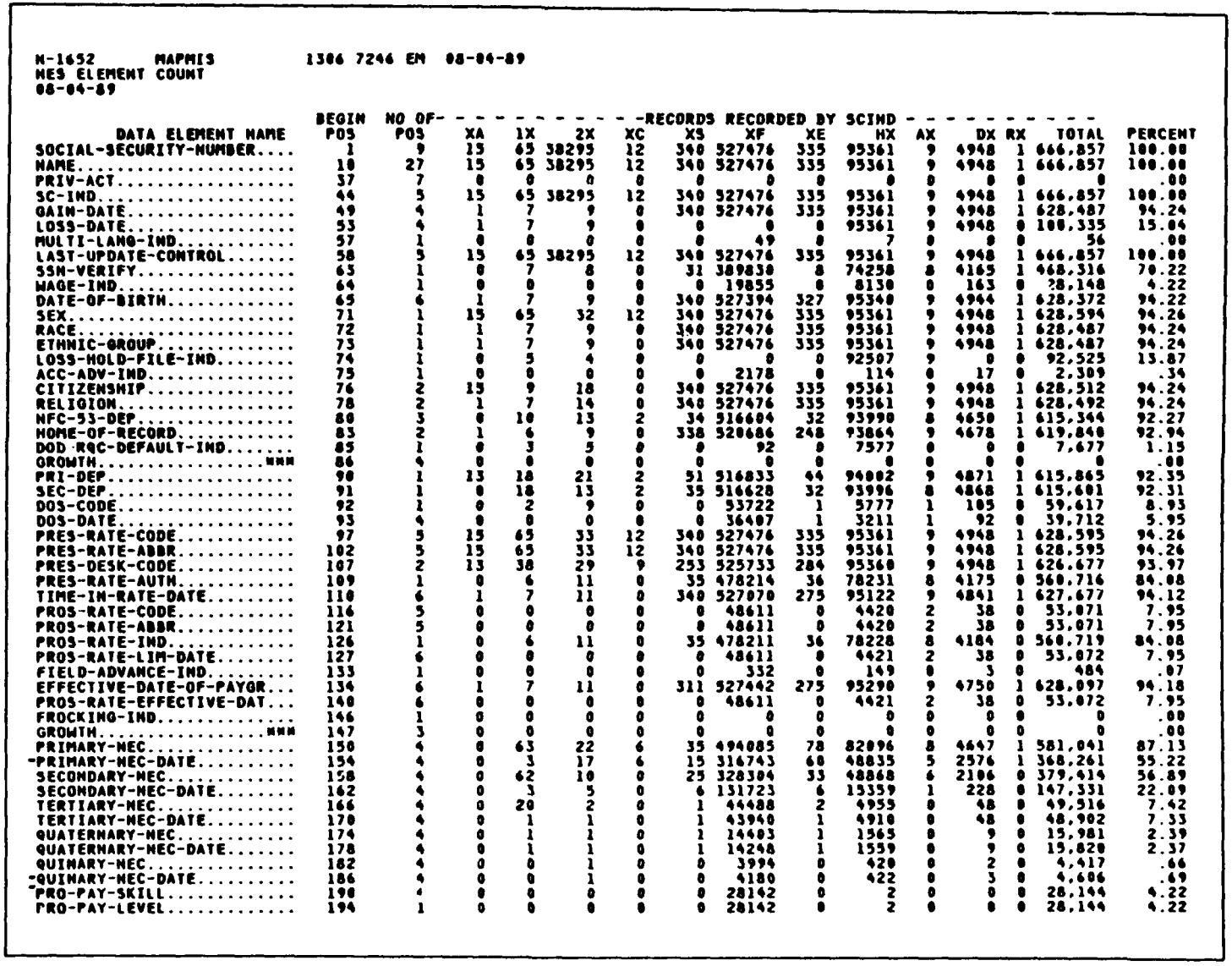
Lastly, Hickman also recommended that incoming NES transactions be maintained on-line for research purposes (Hickman, 1987, p.14). This is already done with transactions in the officer system, and is a valuable tool for identifying trends and resolving processing problems.

5. Methods for Assessing Data Quality

There is no single way to measure the quality of EMF data, and there are no published standards regarding EMF data quality. However, both of these issues, measuring data quality and setting quality standards, are being addressed at all levels of management within OP-16/NMPC-16 and the MPT community. Efforts such as establishing a task force to improve the way timeliness is measured and publishing a policy document about data quality standards, are steps in the right direction. Unfortunately, the problems in both of these areas are complex and not easily resolved. Data quality has a number of different components, some of which can be measured more easily than others. Until data quality can be measured, data quality standards will serve no function as management tools. NMPC-16 regularly measures two of the components of data quality, completeness with reasonable success, and timeliness with increasing success. In addition, the accuracy of selected data elements is measured by making comparisons with other files.

The completeness of EMF data is measured monthly and summarized in the NES Element Count report. This report is useful for assessing the completeness of fields such as sex, place of birth, or term of enlistment, as every individual record should contain an entry. It is less useful for fields such as language ability or school completion date, as absence of data may signify either a lack of the qualification on the sailor's part or an incomplete record in the file. A portion of a sample report appears in Figure 5.

The timeless of inputs to the NES is also measured. This component of quality is more difficult to assess. There are multiple systems putting data into NES and measuring timeliness in a different way. As an outgrowth of this realization, a task force on timeliness was formed. "This issue was introduced at the February 1988 Pay/Personnel Interface meeting with NAVFINCEN [NFC], Code 6 taking the lead . . . This 'task force' was established to standardize the collection, measurements, and reporting of data." (Teter, 1988, p.1) Information regarding the timeliness of one of the systems which provides inputs to NES can be found in Table 5. This information comes from the DMRS, a personnel information collection system, managed by EPMAC in New Orleans. These statistics are based on the date the event occurred compared to the date time group of the naval message in which the event was recorded. In some cases it is appropriate that the event date be in the past, for example a retroactive entry. This skews these statistics slightly. NES is only a part of the Manpower, Personnel, and Training Information System (MAPTIS).



N-1652 MAPMIS 1386 7246 EM 08-04-89
MES ELEMENT COUNT
08-04-89

DATA ELEMENT NAME	BEGIN POS	NO OF POS	RECORDS RECORDED BY SCIND XA	1X	2X	XC	XS	XF	XE	HX	AX	DX	RX	TOTAL	PERCENT
SOCIAL-SECURITY-NUMBER....	1	9	15	65	38295	12	340	527476	335	95361	9	4948	1	666,857	100.00
NAME.....	10	27	15	65	38295	12	340	527476	335	95361	9	4948	1	666,857	100.00
PRIV-ACT.....	37	7	0	0	0	0	0	0	0	0	0	0	0	0	.00
SC-IND.....	44	5	15	65	38295	12	340	527476	335	95361	9	4948	1	666,857	100.00
GAIN-DATE.....	49	4	1	7	9	0	340	527476	335	95361	9	4948	1	628,487	94.24
LOSS-DATE.....	53	4	1	7	9	0	0	0	0	95361	9	4948	0	108,335	15.84
MULTI-LANG-IND.....	57	1	0	0	0	0	0	49	0	7	0	0	0	56	.00
LAST-UPDATE-CONTROL.....	58	5	15	65	38295	12	340	527476	335	95361	9	4948	1	666,857	100.00
SSN-VERIFY.....	63	1	0	7	8	0	31	389838	8	74258	8	4165	1	468,316	70.22
WAGE-IND.....	64	1	0	0	0	0	0	19855	0	8130	0	163	0	28,148	4.22
DATE-OF-BIRTH.....	65	6	1	7	9	0	340	527394	327	95340	9	4944	1	628,372	94.22
SEX.....	71	1	15	65	32	12	340	527476	335	95361	9	4948	1	628,594	94.26
RACE.....	72	1	1	7	9	0	340	527476	335	95361	9	4948	1	628,487	94.24
ETHNIC-GROUP.....	73	1	1	7	9	0	340	527476	335	95361	9	4948	1	628,487	94.24
LOSS-HOLD-FILE-IND.....	74	1	0	5	4	0	0	0	0	92507	9	0	0	92,525	13.87
ACC-ADV-IND.....	75	1	0	0	0	0	0	2178	0	114	0	17	0	2,309	.34
CITIZENSHIP.....	76	2	15	9	18	0	340	527476	335	95361	9	4948	1	628,512	94.24
RELIGION.....	78	2	1	7	14	0	340	527476	335	95361	9	4948	1	628,492	94.24
NFC-53-DEP.....	80	3	0	10	13	2	34	516604	32	93990	8	4650	1	615,344	92.27
HOME-OF-RECORD.....	83	2	1	6	9	0	338	520686	248	93864	9	4678	1	619,848	92.94
DOD RQC-DEFAULT-IND.....	85	1	0	3	5	0	0	92	0	7577	0	0	0	7,677	1.15
GROWTH.....NNN	86	4	0	0	0	0	0	0	0	0	0	0	0	0	.00
PRI-DEP.....	90	1	13	18	21	2	51	516833	44	94002	9	4871	1	615,865	92.35
SEC-DEP.....	91	1	0	18	13	2	35	516628	32	93996	8	4868	1	615,601	92.31
DOS-CODE.....	92	1	0	2	9	0	0	53722	1	5777	1	185	0	59,617	8.93
DOS-DATE.....	93	4	0	0	0	0	0	36407	1	3211	1	92	0	39,712	5.95
PRES-RATE-CODE.....	97	5	15	65	33	12	340	527476	335	95361	9	4948	1	628,595	94.26
PRES-RATE-ABBR.....	102	5	15	65	33	12	340	527476	335	95361	9	4948	1	628,595	94.26
PRES-DESK-CODE.....	107	2	13	38	29	9	253	525733	284	95360	9	4948	1	626,677	93.97
PRES-RATE-AUTH.....	109	1	0	6	11	0	35	478214	36	78231	8	4175	0	560,716	84.08
TIME-IN-RATE-DATE.....	110	6	1	7	11	0	340	527070	275	95122	9	4841	1	627,677	94.12
PROS-RATE-CODE.....	116	5	0	0	0	0	0	48611	0	4420	2	38	0	53,071	7.95
PROS-RATE-ABBR.....	121	5	0	0	0	0	0	48611	0	4420	2	38	0	53,071	7.95
PROS-RATE-IND.....	126	1	0	6	11	0	35	478211	36	78228	8	4184	0	560,719	84.08
PROS-RATE-LIM-DATE.....	127	6	0	0	0	0	0	48611	0	4421	2	38	0	53,072	7.95
FIELD-ADVANCE-IND.....	133	1	0	0	0	0	0	332	0	149	0	3	0	484	.07
EFFECTIVE-DATE-OF-PAYGR...	134	6	1	7	11	0	311	527442	275	95290	9	4750	1	628,097	94.18
PROS-RATE-EFFECTIVE-DAT...	140	6	0	0	0	0	0	48611	0	4421	2	38	0	53,072	7.95
PROCKING-IND.....	146	1	0	0	0	0	0	0	0	0	0	0	0	0	.00
GROWTH.....NNN	147	3	0	0	0	0	0	0	0	0	0	0	0	0	.00
PRIMARY-NEC.....	150	4	0	63	22	6	35	494085	78	82096	8	4647	1	581,041	87.13
PRIMARY-NEC-DATE.....	154	4	0	3	17	6	15	314743	68	48835	5	2576	1	368,261	55.22
SECONDARY-NEC.....	158	4	0	62	10	0	25	328384	33	48868	6	2106	0	379,414	56.89
SECONDARY-NEC-DATE.....	162	4	0	3	5	0	6	131723	6	15359	1	228	0	147,331	22.09
TERTIARY-NEC.....	166	4	0	20	2	0	1	44488	2	4959	0	48	0	49,516	7.42
TERTIARY-NEC-DATE.....	170	4	0	1	1	0	1	43940	1	4910	0	48	0	48,902	7.33
QUATERNARY-NEC.....	174	4	0	1	1	0	1	14483	1	1565	0	9	0	15,981	2.39
QUATERNARY-NEC-DATE.....	178	4	0	1	1	0	1	14248	1	1559	0	9	0	15,820	2.37
QUINARY-NEC.....	182	4	0	0	1	0	0	3994	0	420	0	2	0	4,417	.66
QUINARY-NEC-DATE.....	186	4	0	0	1	0	0	4180	0	422	0	3	0	4,606	.69
PRO-PAY-SKILL.....	190	4	0	0	0	0	0	28142	0	2	0	0	0	28,144	4.22
PRO-PAY-LEVEL.....	194	1	0	0	0	0	0	28142	0	2	0	0	0	28,144	4.22

Figure 5. EMF Data Element Count Report

Table 5. TIMELINESS OF DMRS INPUTS TO MAPTIS (OCTOBER 1988 - SEPTEMBER 1989)

Number of Days	1-5	6-7	8-15	16-30
Number of Transactions	268,518	23,239	20,326	43,816
Percentage of Total	76%	6%	6%	12%

Source: Curran, EPMAC, 1989.

The accuracy of certain NES data is measured in comparison to the data in the JUMPS. This is not a measure of absolute accuracy, but it does document whether the data in NES matches that in JUMPS. The NES/JUMPS disparity standard is one-half of one percent, with a long range goal of zero percent (DCNO MOU, 1989, Tab F). In the past, file disparity statistics were produced and standards were issued for each data element. This is no longer being done, as of the 1989 Memorandum of Understanding (MOU). Since the latest disparity statistics produced were for 1984-1985, they are not current assessments of EMF quality.

There is an organization in the Chief of Naval Operation's staff (OP-16) which provides policy guidance to the MPT IRM community. The branch responsible for Data Resource Management has developed a *MPT IRM Data Quality Guideline*. "This guideline is the first step in promoting a data quality program for MPT corporate data." (DCNO, Data Quality Guideline, 1988, p.2) As a policy document, it must apply to all MPT systems, and can contain only general information about data quality. Therefore, it does not attempt to set quality standards. The NMPC-16 organization, described in detail earlier, is tasked with implementing this policy.

While data quality is and always has been a concern in the MPT IRM organization, user perception of the quality of the data bases has not always been positive. NMPC-16 cannot afford to consider this situation a user problem. If managers fail to use vital data, because they perceive it as unreliable, strategic opportunities may be missed. The Navy's IR Program is founded on the recognition that information plays just such a strategic role in managing Navy business.

As discussed above, user perceptions must be dealt with. Especially now, when user assessments are an important and accepted measure of data quality:

> The three methods generally used to examine the data quality in large files are surveys of end users or clients, samples of entire record files, and samples of active or

current cases. Surveys of end users typically measure "perceptions" of data quality and are fraught with problems of recall, self-report bias, and serious underestimates. (Laudon, 1986, p.6)

However, in absence of a better method surveys may be useful. The bottom line is that NMPC-16 could use another organizational technique for assessing data quality. The picture regarding data quality is not as clear as the department would like it to be. Such oblique statements as "The accuracy of the input data had varied over time," point this out. (Milestone IV System Decision Paper, p.8)

6. Resource Allocation Techniques

Currently, there are no specific techniques used to allocate resources towards improved data quality. Decisions are based on queries from higher echelons, requests from MPT managers, and the gut-feelings of the applications programmers. This is an area where operations research, properly applied, could enhance managers' decisions.

F. THE EMF - TRANSITION TO A DATABASE SYSTEM

In the 1990s, NMPC-16 intends to transition the Officer, Enlisted, Reserve, and Inactive master files, to the Integrated Military Personnel Data Base (IMPDB). This transition affords IR managers and systems designers an opportunity to use not just a data driven strategy, but a quality data strategy, to develop the data base of the future:

The goals of the IMPDB are: to provide a cradle-to-grave view of each Navy member's career; to reflect the official service record:[;] to provide a single, authoritative source for corporate data about Navy members; to be consistent for all functions using the data (standardized); to be organized in the most appropriate way to serve the needs of users of the data; and to be valid and available to users. (Hill, 1988, p.1)

NMPC-16 has completed the first and second iterations of designing the logical data model. The General Functional Requirements (GFR) first published in June of 1987, were revised by Tidewater Consultants in August of 1989. Many areas of IMPDB requirements are covered thoroughly in the GFR. However, the information regarding how improved data quality will be achieved in IMPDB is sparse. The GFR section on IMPDB objectives states that "policies and procedures need to be established and implemented to assure that data quality is maintained," but the only reference to data quality stated that "Stored and transmission data error rates should not exceed industry standards." (Tidewater, 1989, pp.3-1 and 3-4) The *Software Architecture Level I Document* also contains some references to a Management Procedures and Standards System which "will 'house' the rules on procedures and standards with which the MPT commu-

nity must comply." (Software Solutions, 1988, p.6) These include such things as "The overall error rate occurring on pay related transactions will be less than 5%," and "95%" of pay related errors will be corrected in 10 days." (Software Solutions, 1988, p.6) There are also references to a Process Monitoring and Control System which "will monitor the environment, and predefined processes to perform statistical analysis in functional areas." (Software Solutions, 1988, p.6) However, the document itself gives only sample standards and processes, and it later asks "How will the policies and procedures established by these systems be enforced?" (Software Solutions, 1988, p.6) The issues of how to improve data quality and how to measure data quality in IMPDB need to be addressed in more detail now. Management control systems to support this goal need to be defined for future development. Chapter V will explore how NMPC-16 could use improved technology to start designing some of these controls.

G. DATA QUALITY AND THE ROLE OF THE USERS

Up to this point, NMPC-16's role in maintaining the quality of the MPT data has been discussed at length. So as not to leave out an important aspect of data quality control, the following paragraphs survey how the users help maintain data.

1. Data which Impacts Pay

Some of the information contained in the personnel file, affects the pay of the military member. Both the member and the servicing pay office have a vested interest in ensuring that this data which impacts pay is accurate. The Leave and Earnings Statement (LES) is produced monthly. It displays this information (though from a different source - JUMPS) to the member and the payroll clerk. Since these data hurt the sailor in the wallet, they are probably the data users spend the most time and energy trying to maintain.

2. Personnel Data

Every month, the Enlisted Personnel Management Activity produces the Enlisted Distribution Verification Report (EDVR). The EDVR instruction contains the following information, which encourages personnel offices to keep data up-to-date:

> 5. Accuracy of the EDVR - Manning and assignment decisions are based upon information contained in the EDVR. It is extremely important that each activity keep its account up-to-date and accurate by reporting personnel events as they occur and correcting errors when identified. (NAVMILPERSCOM Instruction 1080.1D, 1989, p.2)

The Distribution Support Division (NMPC-47) has many elements of the EMF displayed to detailers in the Personnel Information Module of the Navy Military Per-

sonnel Distribution System. Often, an informal validation of these data is conducted when a detailer is negotiating orders with a constituent.

3. Strategic Data

Data from NES are aggregated and used for such functions as force planning, promotion planning, and accession planning by the Chief of Naval Operations Staff (OPNAV) for MPT. These strategic data can be compared to previous data to identify trends. Sometimes the OPNAV staff can identify error trends through analysis of aggregated data, better than other users, who look at the data an individual at a time.

H. CHAPTER SUMMARY

NMPC-16 is a mature ADP organization, with a long history of managing large and complex information systems. Their efforts to achieve improved data quality have been broad and effective. However, the quality of EMF data cannot be reasonably assessed. It would benefit the users and maintainers of the EMF if data quality could be measured. Users would have more confidence in the data and could set different maintenance priorities for various data elements. Data maintenance resources could then be allocated more effectively. This would also allow the maintainers to develop a comprehensive plan for data quality control. As the EMF transitions to a data base, new opportunities for improvement present themselves. Many of the questions presented in the Data Quality Initiatives framework presented in Chapter II, can be addressed as the data base is designed. At this point data value can be established and drive future data quality enhancement priorities. Chapter IV will propose a method for measuring the quality of EMF data, and Chapter V will propose ways to engineer data quality controls into the new data base and recommend new technologies which can be brought to bear on the data quality problem.

IV. MEASU''NG DATA QUALITY IN THE EMF

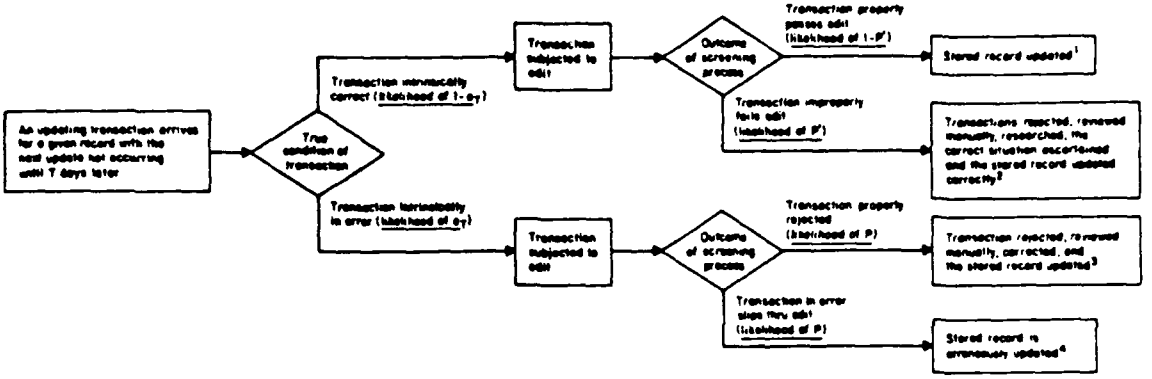
In this chapter, a technique for assessing data quality in the EMF will be presented. The technique is then tested on a small scale. This analysis will serve as a baseline for the Data Management Division, for the Enlisted Research Correction Section, or for other thesis students who may desire to test this technique more fully or devise others which are more appropriate to the NES application.

A. A WAY TO MEASURE DATA QUALITY

NMPC-16 needs a better way to measure the quality of data in the EMF. Haber and associates used linear regression to determine whether good or poor data quality could be related to good or poor input sources (in this case good versus poor reporting was assessed based on quantities of input). This technique was not considered to be appropriate for application to the NES environment, due to the fact that reporting volumes can be correlated with many factors besides accurate reporting. The researchers' assumption that reporting volumes can be correlated with the quality of the data being input, while appropriate for a maintenance system, did not seem reasonable in the case of NES. Ballou and Pazer proposed a model which could produce an expression for error magnitudes in output, but it was only appropriate for applications where all data is numeric.

In 1982, Morey published a paper in which he derived several equations to estimate the stored error rate in a Management Information System (MIS). In deriving these equations, he recognized the importance of the feedback systems concept and the disposition of rejected transactions, on the overall stored MIS error rate. He tested his equations on the leave transaction of the Marine Corps' system for manpower management. In describing the type of system upon which his technique could be used, he said: "The MIS addressed is one where records in a MIS are updated as changes occur to the record, e.g., a manpower planning MIS where changes may relate to a service man's rank or skills." (Morey, 1982, p.337) Since this describes the NES rather well, it appeared feasible that Morey's technique would allow NMPC-16 to quantify the overall quality of the EMF. The estimation technique is based on the relationships between three measures of data quality: the transaction error rate, the intrinsic transaction error rate, and the stored MIS error rate (Morey, 1982, pp.337-338). To describe the proba-

bilities of various dispositions of new transactions, he used a decision tree. This decision tree is presented in Figure 6.



[1] In this contingency, the fraction of the inter transaction interval (0, T) over which the record is incorrect, is the smaller of C_1/T and 1.
[2] In this case, the fraction of (0, T) over which the record is in error is the smaller of $(C_1 + C_2)/T$ and 1.
[3] In this case the fraction of (0, T) over which the record is in error is the smaller of $(C_1 + C_2)/T$ and 1.
[4] In this case the record is in error for the entire interval (0, T).

Figure 6. Morey's Decision Tree

Since it was unrealistic to gather parametric data for all types of transactions processed by NES, only six, from over 130 possible, were identified for data collection. A listing of all the transactions and their purpose is provided in Appendix C. It is important to note that transactions vary greatly in their input format, in their data element contents, and in the associated edit routines through which they are processed.[2]

The transactions were chosen based on several criteria. The first criterion was that failed transactions must be researched, in large part, by an NMPC office. There are some failed transactions which are always returned to the input source for correction, and there would be no simple way to gather statistics on the disposition of these. The second criterion was that the transaction volume and transaction error volume must be great enough that a sample could be obtained within a reasonable time. The last criterion was that there not be known programmatic problems which could create false errors. Transactions were narrowed down to the following: A68, C21, E38, GIB, QC1, 1FL, 1SR, 200, 300, 328, 340, 355, 382, 630, and 798, by scanning a N-1652 Monthly Transaction Totals, Overall Report dated 7 April 1989. A page from this report appears

2 A similarity, every transaction starts with a three character alpha-numeric code to identify its type, followed by the social security number and five characters of the last name, to identify the enlisted member.

in Figure 7. Those transactions listed are the ones with over 100 errors indicated for N-1652 research (the report still says N-1652, as the codes have not been updated since the department reorganized). To decide which of these transaction to collect data on, discussions with various NMPC staff and contractor personnel were conducted. In addition, a transaction not handled by NMPC-16's researchers was selected. Information about the selected transactions appears in Table 6.

Table 6. TRANSACTIONS FOR DATA COLLECTION

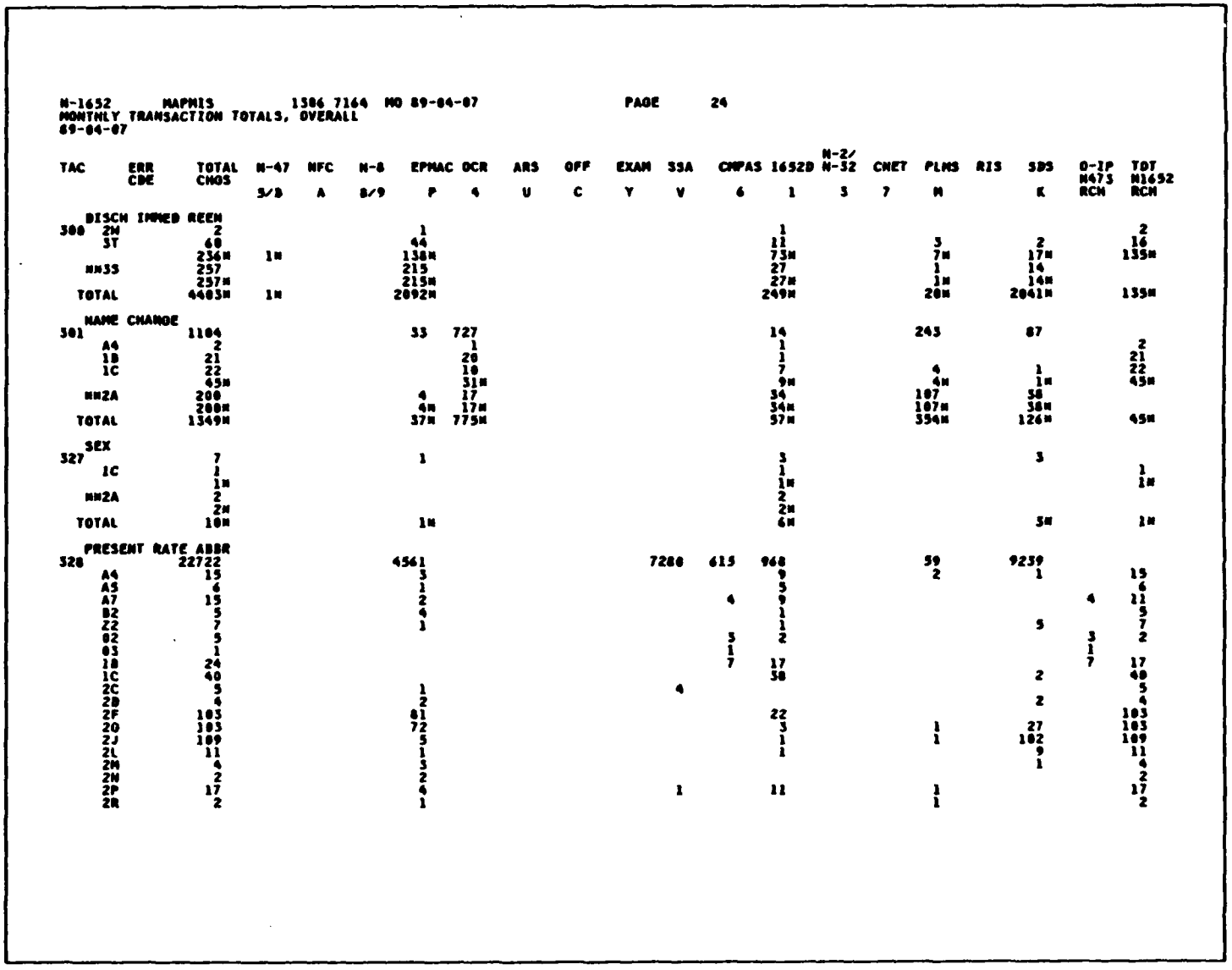
TAC	TAC TITLE	PURPOSE OR DESCRIPTION
A68	Prospective Rate Abbreviation	Update or correct prospective rate.
QC0	Availability	Inform detailers that members are available for immediate assignment.
300	Discharge - Immediate Reenlistment	Process those members who have reenlisted within 24 hours after discharge.
301	Name Change	Change or correct a member's name to agree with official documents.
328	Present Rate Abbreviation	Update present rate.
340	Court Memorandum	Forward to NFC all guilty courts martial findings, all NJP's which affect pay and rate, administrative actions or restoration of above.

Source: Active Duty Enlisted Data Elements Catalog, Appendix A

1. The Parameters in Morey's Formula

An analysis of how each parameter of Morey's estimation formula would be determined was conducted. A detailed description of this analysis is provided in the paragraphs that follow.

First, Morey described the transaction reject rate: "The transaction reject rate [*r* is], i.e., the proportion of incoming transactions which fail, either correctly or incorrectly, the various edits and logical tests used." (1982, p.338) To determine this parameter the N-1652 MAPMIS Monthly Transaction Totals, Overall Report was used. 19 months of data were averaged to determine an overall estimate of *r*. These statistics were taken from 1989 and 1987 reports, mainly because that is what the organization had available for release. The reason only 19 months of data were used is that the reports for June, November, and December 1987, and January and May of 1989 were



N-1652 MAPMIS 1386 7164 MO 89-04-07 PAGE 24
MONTHLY TRANSACTION TOTALS, OVERALL
89-04-07

TAC	ERR CDE	TOTAL CNOS	N-47 5/8	NFC A	N-8 8/9	EPMAC P	OCR 4	ARS U	OFF C	EXAM Y	SSA V	CMPAS 6	1652D 1	N-2/ N-32 3	CNET 7	PLMS M	RIS	SDS K	O-IP N473 RCM	TOT N1652 RCM
DISCH IMMED REEN																				
300	2M	2				1							1							2
	3T	60				44							11			3		2		16
		236M	1M			138M							73M			7M		17M		135M
	MM3S	257				215							27			1		14		
		257M				215M							27M			1M		14M		
	TOTAL	4403M	1M			2092M							249M			20M		2041M		135M
NAME CHANGE																				
301		1104				33	727						14			243		87		
	A4	2					1						1							2
	1B	21					20						1							21
	1C	22					10						7			4		1		22
		45M					31M						9M			4M		1M		45M
	MM2A	200				4	17						34			107		58		
		200M				4M	17M						34M			107M		58M		
	TOTAL	1349M				37M	775M						57M			354M		126M		45M
SEX																				
327		7				1							3					3		
	1C	1											1							1
		1M											1M							1M
	MM2A	2											2							
		2M											2M							
	TOTAL	10M				1M							6M					5M		1M
PRESENT RATE ABBR																				
328		22722				4561					7288	615	968			59		9239		
	A4	15				3							9			2		1		15
	A5	6				1							5							6
	A7	15				2						4	9						4	11
	B2	5				4							1							5
	Z2	7				1							1					5		7
	02	5										3	2						3	2
	03	1										1							1	
	1B	24										7	17						7	17
	1C	40											38					2		40
	2C	5				1					4									5
	2D	4				2												2		4
	2F	103				81							22							103
	2O	103				72							3			1		27		103
	2J	109				5							1			1		102		109
	2L	11				1							1					9		11
	2M	4				3												1		4
	2N	2				2														2
	2P	17				4					1		11			1				17
	2R	2				1										1				2

Figure 7. MAPMIS Monthly Transaction Totals

missing. *Lotus 123* Release 2.01 was used to tally and compute the averages. Averages for two years worth of data were then averaged together for an overall figure. A summary of these transaction reject rates appears in Table 7, and listings of the *Lotus 123* spreadsheets appear in Appendix D.

Table 7. ERROR RATES FOR SELECTED TRANSACTIONS

TRANSACTION	A68	QC0	300	301	328	340
1987	3.10%	12.41%	4.64%	5.15%	13.61%	9.41%
1989	1.20%	8.26%	9.50%	5.33%	22.32%	7.50%
AVERAGE	2.15%	10.34%	7.07%	5.24%	17.97%	8.46%

"***P*** denotes the conditional probability that an erroneous transaction is properly rejected by one of the edits. Hence 1- ***P*** is the probability of the Type I error occurring." (Morey, 1982, p.339) A Type I error is defined as an erroneous transaction that slips through the edits. While Morey does not describe how this can happen, it might be because there are not enough edits, the edits are not stringent enough, or it is impossible for edits to determine that the transaction is in error. ***P*** and 1-***P*** can be estimated by determining the actual status of the transactions that the Enlisted Research Correction Section turns around for five of the transactions. For the QC0 (Availability) transaction, the Enlisted Availability Control Branch recorded the results of their research efforts. Selected members of these sections tallied the status of the rejected transactions they handled on the worksheet at Appendix E. *Lotus 123* spreadsheet summaries of this data appear in Appendix F.

"***P'***denotes the conditional probability that a correct transaction is improperly rejected by one of the edits, thereby delaying proper updating of the record. ***P'*** is the probability of the Type II error occurring." (Morey, 1982, p.339) A Type II error means that a correct transaction is improperly rejected, and 1-***P'*** means that the correct transaction was correctly processed. The probabilities can be measured in the same manner as those above, and the summary data for this parameter also appears in Appendix F. In Table 8, the estimated values for ***P*** and ***P'*** are displayed.

"***T*** denotes the nonnegative random variable representing the time interval or spacing between transactions for a given record of the type being analyzed; these are further assumed to be independent and identically distributed. Candidates for ***T*** might be the exponential, uniform, lognormal random variables or even a constant. Also let

Table 8. VALUES FOR PROBABILITIES: * means that no data was available for this transaction.

TRANSACTION	A68	QC0	300	301	328	340
P	*	.8772	*	.7500	.5000	.9583
P'	*	.1228	*	.2500	.5000	.0417

μ_T denote the mean of the intertransaction times." (Morey, 1982, p.339) Unfortunately, there is no accurate way of measuring how often a particular transaction of a particular type is applied to a particular record in the EMF. For example, a QD0, or a set of orders, will be generated on intervals of the members' tour lengths. This could be every two to five years. A 1FL, a gain to active enlisted strength, occurs only once in the life of a particular record. Ideally, if a transaction file were maintained on-line, it could've been queried to get this information. Since a file like this is maintained for the Officer Master File (OMF), it was reasonable to assume that there might be something similar for the EMF. However, that was not the case. In retrospect, even if there were a transaction file, it would probably not go back enough years to do this type of query, due to the large intertransaction times for many transactions and to the sheer volume of enlisted transactions which are processed. The intertransaction times for these transactions are much greater than those for the transactions Morey studied, however, he did not propose any limits on these times. Since this parameter could not be measured, it was estimated. These estimates appear in Table 9.

Table 9. INTERTRANSACTION TIMES FOR SELECTED TRANSACTIONS

TRANSACTION	A68	QC0	300	301	328	340
DAYS	1460	1080	1460	5475	1460	9000

"C_1 denotes the minimum processing time, measuring the elapsed time from when a transaction is submitted to the system until it updates the record. This occurs for a transaction not rejected by any of the edits." (Morey, 1982, p.339) This would be from one to three days in most cases. In the SDS, transactions are normally up loaded daily. DMRS transactions could take longer if the message system is backlogged with

higher priority communications or if Minimize [3] is imposed. It could also take longer if a problem with format causes the transaction be rejected for research at EPMAC, where field inputs are processed for transmission to NES. DMRS timeliness information appears in Chapter III, however these statistics are based on the transaction's date of occurrence. Therefore, they take into account reporting delays by the activity responsible for personnel accounting and retroactive transactions which constitute a correction. In the future, timeliness will be measured for each increment of processing, and a mean should be readily available. For the purposes of this study 1.5 days was used.

"C_2, a constant, denotes the additional processing time delay, over and above C_1, to manually review and correct transactions which (i) were in error, and (ii) were properly rejected by the edits." (Morey, 1982, p.339) This was measured by recording the number of days which elapsed between when the transaction was first rejected and when it was either corrected or deleted from the error suspense file. The figures derived from the data appear in Table 10. See Appendix F for the raw data.

Table 10. TIME FOR TRANSACTIONS TO BE CORRECTED/DELETED

TRANSACTION	A68	QC0	300	301	328	340
DAYS	2.0	1.2	17.8	3.0	28.0	10.8

"C_3, a constant, denotes the additional processing time over and above C_1 , to manually review and allow to enter into the system any intrinsically correct transactions which were rejected by the edits. It is assumed that the reviewer is able to ascertain the correct situation so that the stored record is updated accurately." (Morey, 1982, p.339) This information was obtained from the Enlisted Research Correction Section. The figures derived from the data appear in Table 11. These data appear in Appendix F.

Table 11. TIME FOR TRANSACTIONS TO BE REINPUT: * means that no data was available for this transaction.

TRANSACTION	A68	QC0	300	301	328	340
DAYS	*	1.9	*	9.0	1.7	2.0

3 Minimize is when message traffic to a particular geographic area is suspended unless it is operationally oriented.

2. Determining the Stored MIS Error Rate

Using the parameters described in the previous section, an estimate for e_T, "The intrinsic transaction error rate, i.e., the proportion of the incoming transactions that are truly in error," can be obtained. (Morey, 1982, p.338) The equation that applies is:

$$\hat{e}_T = \begin{cases} 0 & \text{if } \hat{r} < P' \\ \dfrac{\hat{r} - P'}{P - P'} & \text{if } P' \le \hat{r} \le P \\ 1 & \text{if } \hat{r} > P \end{cases} \qquad (1)$$

However, if $P \le P'$ then the formula does not apply, and the edit should be eliminated, because it is causing the rejection of more correct transactions than erroneous ones (Morey, 1982, p.340). A complete explanation of the derivation of this formula is contained in the appendix of Morey's paper (1982, p.342).

Next Morey defined the stored MIS error record rate, e_M as "the probability that the stored record is in error for any reason. It is defined as the likelihood that a randomly chosen record (for the particular record type of interest) examined at a random point in time, is in error. It includes the situation where a change in the record has occurred but has not been updated in the record." (Morey, 1982, p.338) Notice that this definition takes into account the concept of data volatility, mentioned in Chapter II. Using the equation below, a lower bound for the stored MIS record error rate can be found. This could allow NMPC-16 managers to assess the relative accuracy of the data applied by each transaction. It also could help the Data Management Division to target available resources to enhance data quality more effectively. Morey says:

$$\hat{e}_M \ge \hat{e}_T(1 - P) + [C_1(1 - \hat{e}_T)(1 - P') + (C_1 + C_2)\hat{e}_T P + (C_1 + C_3)(1 - \hat{e}_T)P']/\mu_T \qquad (2)$$

Since the motivation for discussing Morey's paper was to introduce a technique for measuring the overall stored MIS error rate of the EMF, this paragraph explains how that could possibly be achieved. As stated before, equation (2) gives the lower bound for the stored MIS error rate of data elements associated with a particular type of record, or in this case a particular type of transaction. In determining the overall stored MIS error rate of the EMF, this process would have to be repeated for each of the 130 plus transactions, and then that figure would have to be weighted by the data elements contained. Remember, each transaction contains different data elements and a different number of data elements. However, a particular data element may be contained in more than one transaction. That is to say that in regards to data elements the

transactions are not mutually exclusive, but are collectively exhaustive. Perhaps, further research will point out that certain transaction volumes are low enough as to not significantly affect the overall error rate of the EMF, and then they could be excluded to reduce the effort required to obtain the EMF's error rate.

3. The Results of the Analysis

All the values were entered into the equations to determine if an acceptable estimate of the stored MIS error rate could be found. The computed values for e_M varied from .02% to .19%, extremely low error rates. Complete results appear in Appendix G. In two cases, for the QCO and 301 transactions, $\hat{r} < P'$, therefore equation (1) did not apply and the intrinsic transaction error rate, $\hat{e}_T$, was considered to be zero, rather than a computed value which would have been negative. For the 328 transaction, $P \leq P'$. In a case like this, Morey says that the intrinsic transaction error rate can not be estimated by his equation and that the edits should be dropped, because they are not performing a useful function. In addition, for the A68 and 300 transactions, data to determine one of the constants, C_3, was not available. Therefore, the value for C_3 was assumed to be zero. This leaves only one transaction, 340, for which all conditions of Morey's formulas were met, and estimates for all parameters were available. The paragraphs below explain why the data did not meet the conditions of Morey's formulas, and why this analysis was flawed.

The first problem was that insufficient data was collected. The transaction counts provided in the NES F/M statistics, which were used to determine how long to collect data, did not accurately indicate the workload being handled by the Enlisted Research Correction Section. This is because some transactions are now being returned to the input source for correction, through SDS. Until program changes are made on NOCS, there will be no better way to assess what errors the section is actually correcting, versus what erroneous transactions they are just deleting from the error suspense file.

Another problem occurred because the value of $\hat{r}$ was based on almost two complete years of data, while the sample used to derive P' was very small. This leads to making unrealistic comparisons between $\hat{r}$ and P'. Since Morey's article did not explicitly state how he determined the values for various parameters in his equations, it was felt that this approach was sound. The only other alternative would have been to track the disposition of specific transactions, and measure every parameter from the same sample. This would have meant collecting all of the data from the Enlisted Research Correction Section, rather than using some data from reports which are already produced.

It is also important to note that in this analysis, where a parameter could not be determined from available or collected data, it was estimated. Since μ_r, which was an estimated parameter, was very large in comparison to the values Morey used, there is some question as to whether this technique is appropriate for data files where all elements are not regularly updated.

Two final problems influenced the results of this study. During the time that this data was collected, there was some suspicion that duplicate tapes were entered into processing. Confusion as to how to record the disposition of these duplicate transactions complicated matters. In addition, the directions provided to those gathering data were not explicit enough, and the recording sheet was unclear. The disposition of transactions should have simply been categorized in two ways: Correct transaction - reentered as is, or Incorrect transaction - deleted or corrected.

B. RESOURCE ALLOCATION FOR DATA MAINTENANCE

It would be ideal if, in addition to knowing the intrinsic transaction error rates, those rates could be used to allocate resources for data maintenance. Ballou and Kumar.Tayi developed an integer program that does just that. However, the model requires a great deal of data which are not easily attained in the NES environment. A description of the integer program and the difficulties experienced in trying to apply it are explained in the paragraphs which follow.

1. The Integer Program (IP)

Maximize

$$\sum_{i=1}^{n}\sum_{j=1}^{\ell(i)}\left[\sum_{k=1}^{n} P_k N_k e_k p_{ij}(k)\right] x_{ij} \tag{3}$$

subject to

$$x_{ij} = 0 \text{ or } 1, i = 1, \ldots, n; j = 1, \ldots, \ell(i) \tag{4}$$

$$\sum_{j=1}^{\ell(i)} x_{ij} \leq 1, i = 1, \ldots, n \tag{5}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{\ell(i)} x_{ij}(F_{ij} + c_{ij}N_i + \sum_{k=1}^{n} C_{ij}(k)N_k e_k p_{ij}(k)) \leq R \qquad (6)$$

(Ballou, 1989, p.323)

2. Notation for the IP

According to Ballou and Kumar.Tayi:

n number of data sets $S_i = 1, 2, \ldots, n$

P_i cost incurred to the organization for each undetected error in data set i

N_i number of data units in data set i

e_i stored data error rate prior to maintenance for data set i

$p_{ij}(k)$ effectiveness (i.e., ratio of number of errors detected to total number of errors) on data set k ($k = 1, 2, \ldots, n$) of applying maintenance procedure j to data set i, $j = 1, 2, \ldots, \ell(i)$. ($\ell(i)$ is the number of maintenance options available for data set i)

c_{ij} cost per data unit of applying maintenance procedure j to data set i

$C_{ij}(k)$ cost per data unit in data set k of correcting data units identified as deficient as a result of applying procedure j to data set i

F_{ij} fixed cost of maintenance procedure j on data set i

R total resources (in same units as P) available for data quality maintenance

(Ballou, 1989, p.322)

For a more complete discussion of the variables and the model formulation see the paper by Ballou and Kumar.Tayi (1989).

3. Difficulties in using the IP

The IP could not reasonably be applied in the current NES environment. The cost of errors is difficult if not impossible to quantify, and there are over 2700 possible errors. A sample page from the NES error listing is contained in Figure 8. Even if these errors could be grouped into appropriate classes, it would still be a formidable task to quantify each class for each transaction. The effectiveness of various data maintenance techniques has never been measured, and if they had been it would be difficult to use these quantities in an equation that considers them indexes of the same thing. The rea-

son for this is that the effectiveness of transaction error correction would be based on data sets consisting of transactions, just as was done when using Morey's technique. In this case some data elements would be contained in more than one data set. The effectiveness of a file sweep or reconciliation would have to be measured based on data sets which would be made up of data elements. Once a transaction is applied to the file, its structure is lost and cannot readily be recovered. Currently the data sets validated in file sweeps do not exactly match any of the transactions contents. As discussed in Chapter III, it is also difficult to quantify what NMPC-16 is spending for error research and correction, much less for data maintenance as a whole. Therefore, the fixed and variable costs of data maintenance techniques are difficult to ascertain. Even ***R***, the value of the total resources available for data quality maintenance, is not readily available, as some of those functions are being performed by several MPT organizations, including NMPC, EPMAC, and NFC. All of these factors, and undoubtedly others which have yet to be identified, combine to make this task too difficult to undertake within the scope of this study.

C. CHAPTER SUMMARY

Efforts to apply quantitative techniques to the complicated NES environment were not successful, but much was learned in the process. These lessons learned could help in the future, by providing essential background on what data to gather, and how to gather them. When collecting data, directions to data gatherers (who are in most cases junior personnel) need to be precise and clear. For example, the categories for disposition of transactions should have been easier to distinguish. Samples should not be collected during a period of flux, such as when software modifications are being conducted, or when the disposition of errors is being changed. Ambiguities in the application of new techniques should be clarified before data are collected. For example, Morey proposed no upper bound on intertransaction times, however he did state that the record type was one which would receive periodic updates. If certain elements in the record never change, or if others change only after long intervals, does that mean this technique is inappropriate? In addition, collecting one parameter of an equation from historical data, and another from anecdotal data does not seem to be appropriate. The assumption was that it would allow for a more accurate measure of overall transaction error rates, but in some cases, the large sample made this parameter sufficiently low as to invalidate the use of the estimation equation. Finally, it is apparent that if quantitative

```
N-1652 MAPMIS 1306-7262 TAC ERR MSG CODES AS OF JUN 20
TAC     ERROR       DESCRIPTION
300      2B         ENL CODE = 3R; PRES PG MUST BE E1-E3    P
300      2C         EMR SCH/OTH GREATER THAN TRANS TERM     P
300      2D         EMR BR/CL NOT 11,15,32                  P
300      2E         TRANS DT OF OCCUR NOT > EMR EAOS        P
300      2F         EAOS CANNOT EXCEED EREN
300      2G         EMR PEBD INVALID
300      2H         TRANS DOB < 16 YRS PRIOR TO EMR PEBD
300      2J         EAOS EXP-NO EXTENS OR INDIC ON FILE
300      2K         EAOS EXP-EXTENS PRES-MAKE OPERATIVE
300      2N         EMR PEBD INVALID YR-MO-DA
300      2P         DT OF OCC N=00-90 DAYS PRIOR EMR EAOS
300      2Q         EMR BR/CL= 68,78; LOSS CODE NOT=06,46   P
300      2R         EMR BR/CL = 68; ENL CODE NOT= 41 OR 51  P
300      2S         EMR BR/CL = 78; ENL CODE NOT= 11 OR 51  P
300      2T         DT OCC NOT 3MOS-1YR PRIOR TO EMR EAOS
300      2V         EMR BR/CL NOT = 32, 68 OR 78            P
300      2W         STAR, EMR NO. OF ENLISTMENTS MUST BE 1
300      2X         EMR BR/CL= 68,78; TAC BR/CL MUST BE 11  P
300      3K         TRANS PERS LOSS INVALID
300      3M         DT OCC N=00-90 DAYS PRIOR EAOS/EREN
300      3O         EMR NBR ENL N= 1-9, A-F
300      3R         EMR ADSD INVALID
300      3S         **EMR CED > TAC DT OCC
300      3T         EMR BR/CL N= TAC BR/CL                  P
300      3U         EDLN REAS = RRR, RRA, OR QCP
300      4A         DT OF OCC NOT= 00-90 DAYS PRIOR EMR EAOS
301      A1         INVALID SOURCE CODE                     P
301      A4         INVALID NEW NAME                        P
301      02         INVALID SSN RANGE                       P
301      03         INVALID NAME                            P
301      1B         UNMATCHED SSN                           P
301      1C         UNMATCHED NAME                          P
301      2A         **NAME UNCHANGED
327      A1         INVALID SOURCE CODE                     P
327      A4         INVALID SEX CODE                        P
327      02         INVALID SSN RANGE                       P
327      03         INVALID NAME                            P
327      1B         **UNMATCHED SSN                         P
327      1C         UNMATCHED NAME                          P
327      2A         **TAC SEX = EMR SEX
328      A1         INVALID SOURCE CODE
328      A4         INVALID AUTHORITY CODE
328      A5         INVALID PRESENT RATE
328      A6         NEW RATE = BLANK, AUTH N= A,B,C,G
328      A7         INVALID NEW RATE
328      A8         **SRC=6; NEW RATE N= APPRENTICE
328      A9         INVALID EFFECTIVE DATE
328      B2         INVALID TIME-IN-RATE DATE
328      B6         AUTH=0,NEW CODE N= 3600,5000,6000,78000
328      B7         TIR INVALID FOR EFFECTIVE DATE OF TRANS
328      Z2         INVALID DATE OF OCCURRENCE
328      02         INVALID SSN RANGE
328      03         INVALID NAME
328      1B         UNMATCHED SSN
328      1C         UNMATCHED NAME
328      2A         TAC EFF DT > TAC AS OF DATE
328      2B         **DECL OF RATE-NO EMR PROS RATE
```

Figure 8. MAPMIS TAC Error Codes

techniques such as these are to be properly tested, a significant commitment in time and resources will have to be expended in the data gathering phase of the endeavor.

V. DATA QUALITY IN THE EMF OF THE FUTURE

In this chapter, recommendations for providing enhanced data quality in the integrated data base of the future are proposed. These recommendations are explained based on the Data Quality Initiates framework developed in Chapter II, and are compiled and distilled from many researchers' philosophies and efforts. These include: the concept of data value as a decision driver (Varley, 1969), the classification of data quality control techniques based on the SDLC (Brodie, 1980), and the consideration of the probability of being able to maintain a particular data item before deciding to collect it (Davis, 1985).

A. BACKGROUND

The functions of quality control and quality assurance have typically been conducted after a product was designed and produced. This has been true in disciplines as diverse as manufacturing and software development. Fortunately, managers have recognized the importance of ensuring that a quality product is produced, and have started to engineer products with particular quality standards in mind. In this way, either quality is assured, or at least the product is more maintainable. This early emphasis on quality has been used in industry, through initiatives such as TQM, and it has been used in software development, thorough methodologies such as structured analysis and design. At the same time that information science has been making advances in the area of software quality, it has also begun to focus on information as a strategic resource in business. This focus has been fueled by development strategies such as Information Engineering. Now it is time for information systems planners to take data driven strategy a step further, and address data quality throughout the SDLC. This is particularly true for large-scale MISs. Data quality planning can be done by assessing the value of the data that will be collected and the resources which must be expended to maintain the data, and by making an early decision as to whether there are sufficient resources to keep that data at the desired level of quality:

> The relationship between cost, priority, data worth, and error detection and correction procedures should form a set of principles. These principles will enable the system manager or system designer to provide the system users with information that is more accurate and more usable than ever before. (Varley, 1969, p.138)

Admittedly, implementing this strategy is easier said than done, and such problems as being able to do better economic evaluations will have to be resolved. On the plus side, data integrity languages, DD/DS, and Operations Research (OR) are powerful tools, which properly applied, could help improve data quality maintenance.

B. APPLYING THE FRAMEWORK TO THE NEW DATA BASE

In this section, recommendations on how to handle planning for quality control in the data base of the future will be provided. These recommendations are separated into the categories presented in the Data Quality Initiatives framework, based on when in the SDLC they will be used. However, since these initiatives are being considered before systems implementation, to a certain extent, they are all efforts to engineer data quality, in advance of the system's deployment.

1. Engineering in Data Quality

Plans indicate that NMPC-16 will load the IRE with the IMPDB logical data model. Though the data elements have been standardized, more user input should be collected and added to the meta data. A group of users of MPT data should be formed to classify all off the 500 plus data elements, in terms of their value delivered to the MPT organization and the Navy. After looking at just NES data, and trying to assess its value for use in the IP mentioned in Chapter IV, it is apparent that this is a large task. To quantify the value of each data element monetarily would be too difficult. However, it is feasible for the data to be ranked, if a system similar to that used for security risk assessments is employed. In this system, level one data could be those data of strategic importance to the Navy. Level two data would be data of significant strategic importance, and level three data would then be of moderate or minimal importance. Once these classifications are established, associated accuracy and timeliness tolerances could be assigned. These data could then be entered into the IRE, and would have the immediate benefit of providing explicit priorities for data quality control. It would also have a long term benefit. Data could be positioned and accessible if optimization routines using these parameters are developed and implemented. Such prestaging is considered a strategic move in planning for the information technology of tomorrow. The new IRE entries would be similar to those that appear in Figure 9.

2. Software and Hardware Quality Controls

In moving to a data base environment, NMPC-16 will take advantage one of the technological improvements that became widely available in the 1980s. However, managing the quality of the data base will be a new challenge. Along these lines,

ITEM

Pay-Grade

DESCRIPTION

The pay-grade in which the member is currently serving.

PERMISSIBLE VALUES

E1 SR	W1 WO	O1 ENS
E2 SA	W2 WO	O2 LTJG
E3 SN	W3 WO	O3 LT
E4 PO3	W4 WO	O4 LCDR
E5 PO2		O5 CDR
E6 PO1		O6 CAPT
E7 CPO		O7 RADM
E8 SCPO		O8 RADM
E9 MCPO		O9 VADM

VALUE CATEGORY

Level I

ACCURACY TOLERANCE

.01%

TIMELINESS TOLERANCE

30 Days

EDIT/VALIDATION CRITERIA

Changes of only one pay-grade are permitted, unless in association with a disciplinary action or a data base correction.

Others which might apply . . . (could be stated using a data integrity language).

Figure 9. Updated IRE Entry for the Data Element Pay-Grade

NMPC-16 should investigate the use of a data integrity language. These languages are experimental; in fact, in 1986, Date considered them to be hypothetical (1986, p.446). Later, the *DEC* system called *Rdb/VMS* became available, and incorporated some of the controls that Date described in his hypothetical language (1986, p.437). Using a scheme such as Date purports: "Note therefore, that, a transaction can be regarded, not only as a unit of work and a unit of recovery and a unit of concurrency, but also as a unit

of *integrity.*" (1986, p.448) (Date's emphasis) This integrity language could provide needed data quality control at the headquarters and field-input levels.

3. Methods Involving Data Capture

Now that NMPC-16 has moved to on-line input of data at many of the field activities, what further enhancements in data capture can they incorporate into their planning? First, all sites need to be brought up with on-line editing of interactive input. SDS must be fielded overseas, something which has been delayed at least a decade since the original implementation plan. SDS Afloat, to be used on ships, is still under development. The afloat system, like the shore system, will place the data entry, data validation, and even the error detection and correction at the input source. This allows for the most effective data quality control and the most efficient error correction. As SDS Afloat is fielded, redundant systems like DMRS should be eliminated. This standardization will make it easier for the sailors and civilians who report pay/personnel data to learn proper administrative techniques, because there will be only one system with which to cope. At present, SDS and DMRS are not covered in detail at PN "A" School, due to time-constraints and to the systems' redundancy. When one standard system is in place, it will be easier to train sailors and to set policy for common data quality controls. However, it will be a challenge to reduce the multiple reporting systems to a common system and to keep multiple copies of the input edits in synchronization with one another.

A second issue regarding improved data capture is one that NMPC apparently has not considered. For ease of data entry in the field, voice technology offers a quicker and more accurate input method for redundant tasks, which is characteristic of some field inputs. It might be beneficial for the Field Personnel Systems Division, NMPC-167, to investigate use of voice technology in future SDS upgrades.

4. Methods Involving Error Detection

Error detection improvements depend on many of the things previously discussed in this chapter. These are things such as implementation of a data integrity language, cleaning up and synchronizing system's edits, and better training personnel. In addition, in the new data base, a library of queries to detect data base inconsistencies should be developed. These should be run on a recurring basis to identify data that though accurate when originally input, is volatile and no longer valid.

5. Methods Involving Error Correction

It is likely that most error correction techniques will always require human intervention. By their very nature, errors are exceptions in processing, and human be-

ings handle exceptions better than computer systems. However, a policy regarding error correction responsibility needs to be developed. Gradually, more errors are being sent back to the input source for correction, but criteria for identifying when this is appropriate, and when it is not, have not yet been established by NMPC-16.

It would also be valuable if statistics regarding the disposition of errors could be collected in an automated fashion. This would provide better data for a quality estimation technique such as Morey's. This data may have to be collected in at least two systems, NOCS for in-house corrections, and SDS for those errors turned-around in the field.

6. Methods for Assessing Data Quality

NMPC-16 needs to dedicate some resources to experimenting with ways to assess data quality in NES and in its other corporate data bases, which will eventually be integrated. Morey and others have proved that estimation techniques for measuring data quality can be developed. While this study's attempt to apply that technique was not particularly successful, further investigation into quantifying data quality in the NES and other corporate systems is warranted. Assessment is a component vital to effective management of many aspects of information technology.

7. Resource Allocation Techniques

Chapter IV briefly mentioned an IP to allocate resources to data quality maintenance techniques. OR is being used in many disciplines, to consider decision parameters too complex for the human decision maker to manage. Faculty and students of the OR department at the Naval Postgraduate School are interested in Navy applications for modeling. NMPC-16 should team up with OR professors and thesis students to see if a resource allocation model more appropriate to NES or its future environment can be developed. The model proposed by Ballou and Kumar.Tayi appeared appropriate for application to the NES data maintenance environment, but required too many inputs to make its use feasible. An OR student many be able to take some of the parameters recommended for inclusion in the meta data of the new integrated data base and formulate a model.

8. Fitting the Recommendations Together

How might NMPC-16 use these recommendations together to develop a program for data quality in the future? Figure 10 provides an illustration of how these initiatives can fit together, creating an environment where data quality is actively managed.

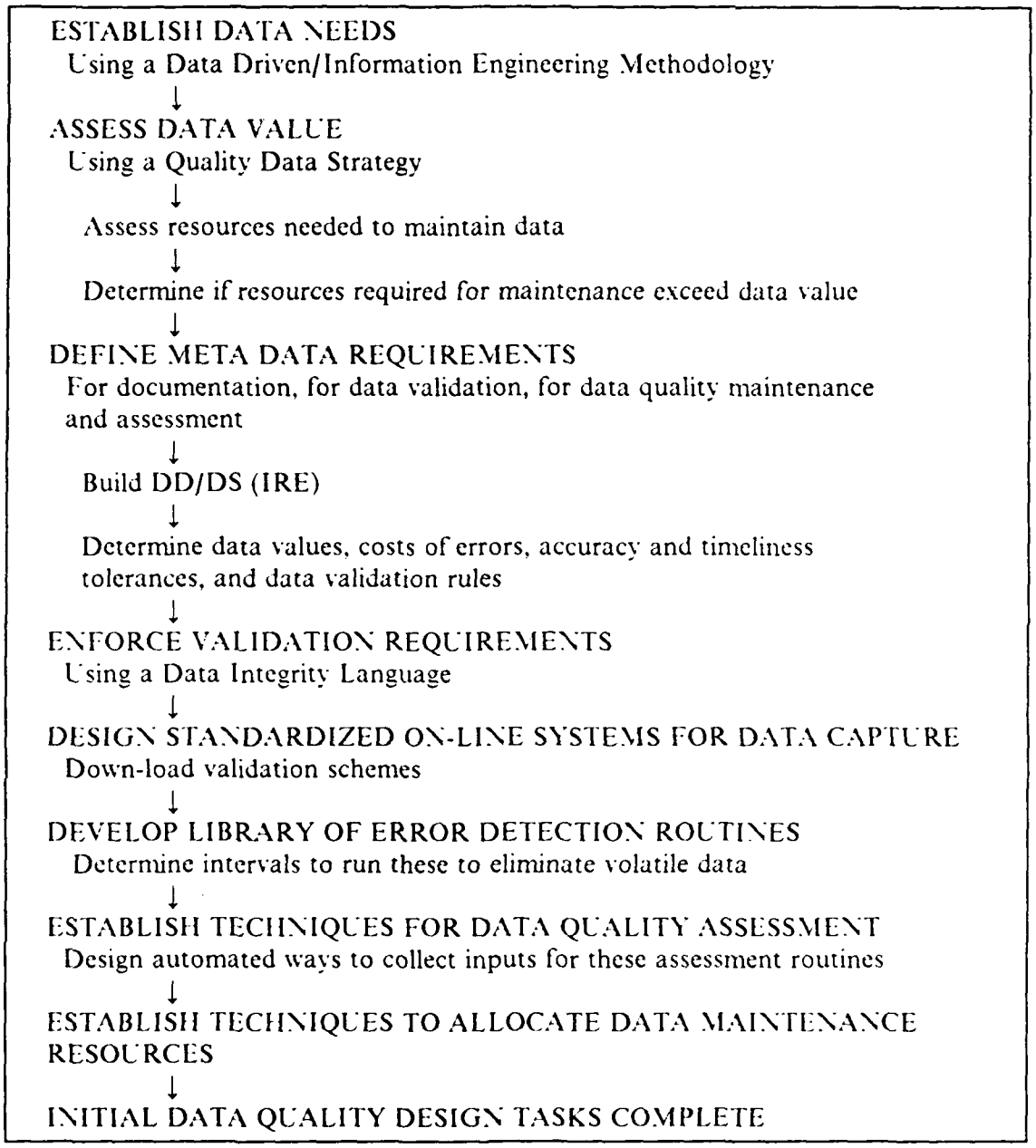


Figure 10. Managing the Quality of the Personnel Data Base

C. CHAPTER SUMMARY

This chapter has proposed that NMPC-16 dedicate resources to exploring several new technologies which could contribute to enhanced data quality in the future. It is appropriate that these be considered now, as the GFR for the integrated personnel data base is being refined and SDS Afloat is being developed. The technologies which

NMPC-16 should investigate include: defining data values with data integrity languages, capturing data with voice technology, and allocating data maintenance resources with OR. Bringing these technologies into the mainstream will require that the MPT community assess the benefits to be gained from accurate information, and that the MPT IRM managers convince resource sponsors that capturing these benefits is a high priority.

Some of the initiatives proposed here may be too ambitious, too costly, or for one reason or another may not fit within the broader strategic goals of the of the MPT or MPT IRM communities. However, this analysis sought to take a new look at an old problem, and perhaps germinate an idea which will mature into a valuable program. In this ambitious no-holds-barred approach to solving the data quality problem, an attempt was made to fit recommendations into programs already articulated, however resource constraints were not considered. Admittedly, MPT IRM managers do not have this freedom, so some of the ideas presented here may seem a little far-fetched, however: "At resource allocation time, the difference between an effective strategic initiative and a harebrained scheme is razor thin. Only after the passage of money and time is the answer obvious." (Cash, 1988, p.145)

VI. CONCLUSIONS

Solving the data quality problem within the NES is not easy for three reasons. First, NMPC has already effected a broad range of measures which have positively influenced the EMF's data quality. Therefore, common concerns like improving data capture through on-line systems have already been identified, and improvements have been initiated. Second, a study of current research and new technology did not reveal any generic data quality improvement schemes, which could be easily applied to NES. Information professionals have not devoted much attention to the data quality issue, until recent years. In the past, maintenance of data quality has been restricted to correcting rejected inputs. To complicate matters, data maintenance problems are unique to each environment. Also, assessment of data quality has typically been left to auditors. However, NMPC-16's future Chief Information Officer of cannot affort to dismiss an issue which is one of the users' chief concerns. Third, NES must function in a complex environment. The data are input at diverse locations, the systems' interfaces are extensive, the users' functional requirements are elaborate, and proper management control is difficult to achieve. Davis says that "The ability of an organization to maintain data quality depends on both organizational factors and data factors." (1985, p.611) These are:

1. Length of error effect cycle
2. Regularity of measurement
3. User-provider link
4. Provider data discipline
5. Ease of verification

(Davis, 1985, p.611)

In the paragraphs that follow a brief analysis of each of Davis's criteria will point out why there are no magic solutions to the data quality problem.

The length of the error effect cycle in NES is varied. For pay-related data items the error affects the sailor almost immediately, or at least within 15 days. Therefore, pay-related errors are much more likely to be corrected. However, other errors may remain buried in the file, impacting only statistical analyses and MPT policy formulation. These types of errors are more insidious and have a much greater potential for affecting the

health of the file over the long term. While it is true that pay-related errors can cost the government money if they are not identified, in most cases, improper payments are later recouped. It is more difficult to assess the impact of errors on concerns such as strength and accession planning.

Another factor which affects the MPT IRM community's ability to control the quality of NES data, is the fact that its accuracy is not regularly measured. The completeness of the data is reported, and timeliness will be measured with more accuracy in the future, however a report on overall data quality is not produced on a recurring basis. Assessments are made only when certain high visibility issues surface, for example, most recently the quality of eligibility data for the G.I. Bill.

The user-provider link in NES is not very strong. The primary users of the data are headquarters organizations. The data providers do not necessarily feel a responsibility to these organizations, and sometimes because of the extensive requirements levied by headquarters on the already over-tasked field activities, an adversarial relationship exists. This relationship is improving, because SDS provides a tighter link between headquarters and the field. Now headquarters' data can be reconciled with field data in an automated manner, and used to produce reports for local management control. Examples of some of these reports include projected rotation date reports and end of active obligated service reports. The user-provider link has been strengthened for those activities that have SDS.

The provider data discipline is a concern, because so many different sites input data into NES. It is difficult to oversee discipline in all the activities which must report personnel and pay data. For the most part, the shore establishment is tightly managed, Personnel Support Detachments (PSD) exist expressly for the purpose of providing personnel/pay support. However, even in PSDs there are competing concerns, such as pay days, temporary duty processing, transfers, and advancement examination cycles. It seems logical to speculate that these pressures only multiply in a operational environment, where accurate reporting of data to headquarters is a minor concern. Training each of the individuals who will have the potential to impact the quality of NES data is a formidable task at best.

Finally, and perhaps the most significant problem in achieving better data quality in the EMF is the difficulty in verifying it. The only way to accurately validate the data is to go to the individual source documents and check them. This can be done by using the paper service record in the field or the micro-fiche record at headquarters. Both of

these validation procedures require manual intervention, and are time-consuming at best.

Evaluating NES with respect to Davis's factors points out just how difficult it is to identify a way to improve data quality, which will make a significant difference and will be cost-effective. However, simply correcting errors does not represent adequate quality control at the headquarters level. While NMPC-16 has introduced initiatives which co-incidentally improved data quality in the NES and other systems they manage, these initiatives were not undertaken as part of a data quality control plan.

A. IMPROVING THE DATA MANAGEMENT ORGANIZATION

The reorganization of 1988 has not achieved the changes needed to improve data quality. This reorganization sought to make NMPC-16 more data-oriented, and even established a separate Data Management Division. Though the reorganization was official in April of 1988, many of the division's management positions were filled with acting directors or left vacant. While the reasons for this were beyond the immediate control of local managers, it has hurt the development of the Data Management Division none the less.

The Head of the Data Management Division is a Navy Captain who is an Acting Director. He is double-hatted, responsible for both the policy and implementation sides of data quality. This Acting director should be left in charge of policy setting; it is valuable to have blue-jacket influence in that area. However, a computer specialist should be hired to run the NMPC/Implementation side. The person hired to fill this position should have a background similar to that usually associated with an Electronic Data Processing (EDP) Auditor. This person could then be tasked with working the more technical data quality issues. There are several reasons why having someone with an EDP auditing background would enhance the effectiveness of this division:

- By trade an auditor is oriented toward assessment, they are trained to place value on the role that assessment plays in proper management. EDP auditing is a specialized field which attempts to measure the effectiveness and accuracy of various aspects of computer systems.
- In addition, an auditor's second responsibility is to make recommendations for improvement to systems they review. This fits in with the primary mission of the data management division, which is to identify ways to maintain and improve the quality of the corporate data bases.
- Much of the current research concerning data quality has been conducted by auditors and has been published in journals for the auditing profession. In many cases, it appears that members of this profession are more familiar with the concerns of measuring and maintaining data quality than ADP professionals.

Since there is no specific job series for EDP Auditing in the civil service, the position description would have to be written from the Position Classification Standards for both the Computer Specialist Series GS-334, and the Auditing Series GS-511. It would be best to classify the position as a GS-334 overall, because it would better fit the structure of the organization. In addition, positions which encompass primarily auditing tasks are controlled by the Office of Personnel Management, and cannot be filled locally. The position should be established at the appropriate level, and titled Computer Specialist, rather than programmer, analyst, or technician. The position will carry with it significant responsibility. The priorities of the division will be based in large part on what the individual who fills this position views as concerns, both with managing data quality in the current systems, as with planning for enhanced data quality in future systems. This employee will have to possess "Broad knowledge of data processing methods, equipment types, systems, applications, and management principles. . . ." (Position Classification Standard for Computer Specialist Series, 1980, p.120) The individual must have a comprehensive knowledge of technologies which might impact on data quality. When wearing the auditing hat this manager "Develops, coordinates, and issues technical audit guidelines and instructions for the inspection of operation and support programs and systems usually at the installation level." (Position Classification Standard for Auditing Series, GS-511, 1982, p.74) "The auditor must justify critical findings and sell recommendations improving the efficiency and effectiveness of agency programs." (1982, p.88)

The Data Quality Program Section is another area of concern. A manager for this section was not installed until November of 1989, over a year after the reorganization took effect. Up until that time there was no comprehensive data quality improvement plan, in fact, there were basically no plans at all. Now that there is some stewardship in this section, perhaps a comprehensive set of internal controls can be established for maintaining an essential corporate resource, personnel data. To develop these internal controls, the Data Quality Program Section should conduct a review of the data maintenance activities performed throughout the organization. Particular emphasis should be placed on those still being done in the Corporate Data Systems Division. Some of these activities might best be handled elsewhere. The reports and queries run by contract personnel in that division, should probably be handled by NMPC-1642C or NMPC-1641E. In any event, the bottom line is that the only data analysis work currently being done for the EMF appears to be inadequate, and is being done in the wrong division. Steps are being taken to resolve this issue, but it will definitely benefit the new manager of the Data Quality Program section to review the findings of the audit reports

discussed earlier and the recommendations of this thesis when restructuring the priorities of the division.

Another concern with the new organization relates to the responsibilities of the Customer Support Division (NMPC-163). It would be valuable if, until another assessment technique is developed, this division could survey major users with regards to data quality. This could provide the Data Management Division with feedback to help set priorities among quality improvement efforts. Transferring error correction assistance for field activities and individuals to the Customer Support Division should also be revisited. This was previously suggested in several management studies. Perhaps a field liaison office could be established within the Customer Support Division. An alternative to this would be to move these functions to the Field Personnel Systems Division for SDS customers and to EPMAC for DMRS customers. This transferring of error correction assistance for field activities and individuals to other divisions will be beneficial only if it doesn't require a duplication of the expertise already available in the Data Management Division.

B. ALLOCATING DATA MAINTENANCE RESOURCES

Some of the methods currently used for data maintenance are too costly. NMPC-16 prints Officer Data Cards (ODC) once a year and on demand, for an officer to verify the personnel data contained in the Officer Master File (OMF). In the past, production of a similar Enlisted Data Card has been considered and rejected. Producing ODCs once a year and on demand is inefficient. Reducing ODC printings might free enough resources to implement a verification scheme for enlisted data. Using this new scheme, a verification record could be sent to activities via SDS. For E1 - E6, verifications could be conducted when the member reenlists. For E7 - E9 and officers, verifications could be completed six months prior to any selection board action on behalf of the member. This would save the funds used for ODC mailings and forms production, and could reduce the ODC correction workload. The correction workload would be reduced, because ODC verification could be coordinated by PSDs and activities providing personnel support, so officers would not be as likely to send in corrections to accurate data that they just don't understand. These resources could be reprogrammed into developing logic for producing the validation runs. This would have the added benefit of providing a common means for verifying OMF and EMF data in advance of their integration. The Officer Distribution Control Report (ODCR) and the EDVR in their current form will also be unnecessary in the future. When SDS is fully deployed, the purposes of the

ODCR/EDVR should be reexamined and unnecessary commitments of resources should be eliminated. Any changes which are considered should be evaluated with the impact of the integrated data base in mind.

More effort should be put into using programs like what one NMPC staff member called "bubble-up". This program compares names and social security numbers of officer and enlisted personnel from the tape that produces the navy locator, and prints any pairs which are exact or close matches. These pairs of records can then be researched and redundancies between the files eliminated. Similar programs could be developed to check information in the file against current tables or to identify data which is no longer valid. Routine queries such as these should be used more often, because they contribute to improved accuracy and are efficient.

C. BETTER TOOLS FOR MANAGING DATA QUALITY

NMPC also needs to develop better tools for managing data quality. For example, NMPC-16 has discussed keeping on-line NES statistics in order to better analyze the reports produced in the daily updates. This is a good idea, which should be implemented immediately, and added as a requirement to the GFR for IMPDB. A thesis student could probably design the data base and load it with as many old statistical reports as are available. In addition, another thesis student could interview experienced NES programmers and users and write some rules for an Expert System to analyze the report data.

Standard Operating Procedures (SOP) should be developed for error researchers. These could go a long way to help train new workers or to serve as a desk guide for experienced researchers. This was recommended by auditors and is reiterated here. Much can be learned by documenting procedures and solidifying policy. Research priorities established in the IRE could be reinforced in this SOP. In addition, responsibilities for error correction of particular transactions needs to be reevaluated. Researchers in the Enlisted Research Correction Section are deleting many of the transactions which appear in NOCS. This is because they are being sent back to field activities for correction via SDS. This needs to be resolved in software, so controls over deletions in NOCS can be established and enforced. Benoit suggests that "Adding or deleting entire records from the error [rotating] file is [should be] avoided . . ." (Benoit, 1979, p.27)

Another audit recommendation was that an on-line transaction summary be kept for NES just as it is for the Officer Personnel Information System. This on-line information

enhanced the quality of error research for the officer system, and could have the same effect on the enlisted system.

The IRE needs to become an active system. Simply storing meta data only increases awareness and articulates standards. It does not enforce MAPTIS-wide standardization. If validation routines could be generated for all systems from the IRE or a DD/DS system, this would greatly improve data quality.

Finally, a way to assess data quality and to allocate resources for data maintenance needs to be developed and institutionalized. Once a methodology is in place, it can always be expanded or refined. The methodology can serve as a baseline to help managers in decision making and to rationalize the allocation of resources to data maintenance. For example, using Morey's method, if it is determined that e_r is too large then:

> Improving the quality of the incoming transactions, i.e., reducing e_r presumably could be accomplished by more training of the preparers of the transactions, use of optical character recognition (OCR) equipment, more emphasis on the care exercised in preparing transactions, etc. (Morey, 1982, p.341)

This quote is not presented to advocate the use of OCR or any other improvement technique, it just serves to point out that by evaluating parameters of data quality, resources can be targeted at problem areas. Several more examples of this appear below:

- "Reducing the cycle time for processing of transactions, i.e., reducing C_1, this could be accomplished by batching more frequently or use of an on-line operation." (Morey, 1982, p.341)
- "Reduction of the time required for manual review, researching, and correction of rejected transactions, i.e., reduction of $C2$ and $C3$. This depends upon more and/or better trained clerks, as well as improving their access to the historical records or individuals involved. This is exactly the thrust of the Navy's new PASS system mentioned earlier where there is to be one single location for each Navy Person, handling all payroll, re-enlistment, separation, vacation, etc. issues. This improved interface will facilitate the researching and correction of rejected transactions." (Morey, 1982, p.341)
- "A tightening of the edits to reduce the frequency of Type I and Type II errors, i.e., to increase P and reduce P'. This requires more precision in the screens used and requires a careful analysis of the relative advantages and disadvantages of deleting or adding edits." (Morey, 1982, p.341)

However, in moving to this form of management by assessment, cost must be a major consideration. The following observation was made in regards to assessing software quality: "The major deterrent to incorporating a measurement program is cost. If the cost outweighs the benefits, the measurement process is not worth pursuing." (Valett,

1989, p.137) NMPC-16 will have to weigh the costs and benefits before deciding to develop techniques for data quality assessment.

D. SUMMARY

Data quality is difficult to define and difficult to quantify, and improved data quality is difficult to obtain. Many initiatives can contribute to improved data quality, but these should be coordinated in an overall program. Developing a comprehensive data quality program is not a simple task. In order to improve their control of data maintenance, NMPC-16 needs to survey current initiatives which contribute to data quality, and decide what future environment is desired. Once that is decided, these goals should be articulated in a data quality plan which could include many of the initiatives suggested in Chapters V and VI of this study. Overall, NMPC-16 has been very proactive in striving to provide accurate and timely data to the MPT community. The initiatives suggested in this thesis are merely refinements to a program which is already on the right track. It would have been easier to make suggestions for improving data quality, if initiatives such as SDS and the IRE were not already started. However, the survey of Data Quality Initiatives conducted in this study pointed out some weak spots, and can provide a valuable assessment tool for other organizations evaluating their data maintenance environment. Since the data-oriented reorganization is less than two years old, the SDS Afloat and the IMPDB are under development, and some parts of the NES are scheduled for recoding, now is an ideal time for NMPC-16 to take the lead in addressing data quality control as an IRM concern.

APPENDIX A. LIST OF ACRONYMS

ADP - Automated Data Processing

CDC - Consolidated Data Center

CIRMP - Component Information Resources Management Plan

CNO - Chief of Naval Operation

CNP - Chief of Naval Personnel

DBMS - Data Base Management System

DD/DS - Data Dictionary/Directory System

DDP - Distributed Data Processing

DE - Data Element

DMRS - Diary Message Reporting System

DON - Department of the Navy

EAM - Electronic Auditing Machinery

EDP - Electronic Data Processing

EDVR - Enlisted Distribution Verification Report

EMF - Enlisted Master File

EPMAC - Enlisted Personnel Management Center

F/M - File Maintenance

GFR - General Functional Requirements

GS - General Schedule

IBA - Information Benefit Analysis

IMPDB - Integrated Military Personnel Data Base

I/O - Input/Output

IP - Integer Program

IR - Information Resource

IRE - Information Resources Encyclopedia

IRM - Information Resources Management

IRSTRATPLAN - Information and Related Resources Strategic Plan

JUMPS - Joint Uniform Military Pay System

LDM - Logical Data Model

LES - Leave and Earnings Statement

MIS - Management Information System

MAPTIS - Manpower, Personnel, and Training Information System
MOU - Memorandum of Understanding
MPT - Manpower, Personnel, and Training
NEC - Navy Enlisted Classification Code
NES - Navy Enlisted System
NFC - Navy Finance Center
NMPC - Naval Military Personnel Command
NMPC- 16 - Total Force Information Systems Management Department
NMPC-163 - Customer Support Division
NMPC-164 - Data Management Division
NMPC-1641 - Corporate Data Maintenance Branch
NMPC-1641E - Enlisted Research Correction Section
NMPC-1642 - Data Implementation Branch
NMPC-1642C - Data Quality Program Section
NMPC-165 - Corporate Data Systems Division
NMPC-166 - Field Personnel Systems Division
NMPC-167 - Technology Support Division
NOCS - NES On-line Correction System
NRA - Navy Recruit Accession
OCR - Optical Character Recognition
ODC - Officer Data Card
ODCR - Officer Distribution Control Report
OMF - Officer Master File
OP-16 - Total Force Information Resources and Systems Management Division
OR - Operations Research
OSD - Office of the Secretary of Defense
PSD - Personnel Support Detachment
RAS - Resource Accounting System
SDLC - Systems Development Life Cycle
SDS - Source Data System
SECNAV - Secretary of the Navy
SOP - Standard Operating Procedures
TAC - Transaction
TQM - Total Quality Management

APPENDIX B. SUMMARY OF NES TRANSACTION STATISTICS

TRANSACTIONS AND ERROR RATES PER MONTH

YEAR AND MONTH OF REPORT	TOTAL TRANSACTIONS PROCESSED	OVERALL TAC ERROR RATE	TAC'S INPUT BY N1652	TAC'S FOR N1652 RESEARCH
8512	764816	21	4794	25290
8601	502312	24.8	2881	5592
8602	725094	19.9	6538	16976
8604	560024	20.5	12686	25951
8606	763363	14.9	17953	21814
8607	631439	12.9	5892	14095
8608	576737	10.5	13612	10573
8609	561344	8.7	5182	6826
8610	508379	11	5282	8598
8611	719261	13	6767	12932
8612	568107	13.2	14025	11036
8702	703976	19.5	6888	7940
8703	533534	22.8	3944	6925
8704	695772	6.9	4991	7965
8706	563128	7.5	5109	17966
8707	645609	12.5	15519	54530
8708	589549	10.2	7110	36926
8709	631230	19.7	35478	51870
8710	609015	7.7	10529	27493
8711	532247	10.5	11755	34350
8712	569784	8.4	19529	27021
8802	728831	7.5	40862	31924
8803	505411	7.5	10191	17000
8809	649806	6.6	5297	22616
8902	452770	6.2	10640	13455
8903	609843	6.1	112090	16145
8904	614925	4.9	41856	14527
8905	611708	6.3	92849	9635
8907	692946	6	27990	20133
8908	595520	8	34375	22281
8909	698265	4.6	78487	14884
8910	556664	6.4	13350	21480
AVERAGES	614732	11.4	21389	19898
SUMS	19671409		684451	636749
YEARLY AVG	7376778		256669	238781

APPENDIX C. TABLE OF NES TRANSACTIONS (TAC)

TAC	DESCRIPTION OR PURPOSE
A62	Align previous rating (enlisted rank and specialty) with that in the member's service record.
A68	Update or correct prospective rate.
B01	Input or change a code indicating what degree commissioning program member is participating in.
B29	Align the basic test battery scores with those in the member's service record.
B48	Input or change a language ability.
B55	Input, change, or correct security data.
B77	Change the six year obligator code for those enrolled in an Advanced Electronic Field.
CAC	Build a skeleton record on a member who enlisted in the Delayed Entry Program.
CAD	Delete members who have enlisted in the CACHE program and not reported or lost those reported erroneously by USAREC.
C03	Align reserve contract extension with that in member's service record.
C04	Input an estimated date of loss to the Navy and a reason for loss.
C21	Align type of enlistment and type of acquisition data with that in the member's service record.
C24	Align Military Obligation Designator data with that in the members' service record.
C25	Correct the number of enlistments and the member's Branch and Class of service.
C26	Input the reason member is retained on file beyond the end of their active obligated service.
C32	Change the member's date of birth.
C39	Change or correct the member's citizenship.
C40	Align religion with that in member's service record.
C41	Align home of record data with that in member's service record.
C43	Correct place of birth.
C70	Correct AFQT score.
DIS	Update type and date of last discharge.
EMR	Provide EMF record for research.
ETP	Update special program or ship data.
E07	Update the career history fields.
E38	Input or correct school history.
E45	Change or correct any one of five entries for school history.
E77	Align ethnic group designation with that in the member's service record.
E85	Correct the special program code.
E89	Input or correct the program availability code.
E99	Correct the education field.
FBK	Report action taken by NMPC on errors from NFC.
FEE	Correct DOD AFEES code.
GIB	Input or change G.I. Bill Eligibility data.
MOB	Enter mobilization gains.
NEA	Change Navy Enlisted Classifications (NEC).
NEB	Add one valid NEC earned through on the job training.
NEF	Add or delete an NEC earned through school.
NFC	Permits NFC to report errors on pay transactions.
NSP	Correct the special program indicator.
OFE	Change the Success Chances for Recruits Entering the Navy Code.
PA1	Purpose is to input the date Privacy Act data was contested.
PA2	Delete the date Privacy Act data was contested.
QAB	Enter the test scores from the Special Assignment Battery.
QAP	Enter the Recruit Assistance Program data.
QA2	Change, delete, or add detailers information.

TAC	DESCRIPTION OR PURPOSE
QA3	Change the distribution NEC.
QA4	Update data concerning overseas assignment requests.
QA7	Remove a flag before detailer makes an assignment.
QA8	Input a flag after a detailer makes an assignment.
QB3	Delete an availability.
QB4	Change special case codes.
QB5	Change the Nuclear Field Indicator.
QB6	Obtain an assignment document.
QCT	Field to allow command turbulence to be monitored.
QC0	Inform detailers that members are available for immediate assignment.
QC1	Inform detailers that students are available for immediate assignment.
QDO	Establish a set of transfer orders.
QFR	Provide prospective fleet reserve (retirement from active duty) information.
QFO	Cancel a set of transfer orders.
QG0	Update GUARD reenlistment data.
QH0	Reprint a set of orders.
QI0	Cancel a previously recorded tour extension.
QI3	Change or correct the Assigned Rate.
QI4	Change the on board orders cost data or permanent change of station data.
QI6	Change the projected rotation date and projected rotation reason.
QI7	Report, change, or correct E5 through E9 evaluations.
QI8	Correct height and weight.
QI9	Print or delete and evaluation history record.
QJ0	To enter a tour extension.
QK0	Enter a prospective gain and distribution data.
QL0	Modify previously recorded duty preference information.
QM0	Modify previously recorded assignment data.
QND	Enter NMPC Code number of the detailing office responsible for member s assignment.
QP0	Enter duty preferences as submitted by members.
QRB	Record selective reenlistment bonus data.
QSD	Identify personnel who require special consideration for assignment.
QS3	Enter test, education , advanced electronic field, and lateral conversion data.
QYO	Enter or change the six year obligation indicator.
Q45	Enter the military spouse identifier.
Q46	Change the sea duty commencement date.
Q51	Change the special category code.
Q56	Change the shore duty commencement date.
Q99	Delete a complete record (of a certain type).
SSV	Input the Social Security Administration verify key and wage status indicator.
TEM	Used to correct data elements as necessary.
TIR	Update time in rate.
TSC	Correct term status code.
WAT	Add or remove a special interest code.
050	Report members who are unauthorized absentees, deserters, or in civilian custody.
051	Process NEC data which adjusts EMF elements due to member s unauthorized absence.
1FL	Process members who are accessions to enlisted strength via AFEES.
1SR	Process active duty members to full strength.
1XX	Establish a skeleton master record.
198	Cancel a erroneously applied strength loss and restore to master file.
200	Gain member on board an activity.
298	Cancel an activity loss processed in error, and reinstate member on board.
300	Process those members who have reenlisted within 24 hours after discharge.
301	Change or correct a member s name to agree with official documents.

TAC	DESCRIPTION OR PURPOSE
327	Correct the member's sex to agree with official records.
328	Update present rate.
330	Process or change Total Obligated Submarine Service data.
331	Correct or modify submarine pay data.
333	Report or change special qualification not identifiable by NEC or rate.
334	Change or correct member's branch and class of service to agree with the service record.
336	Align dependency status with NFC and service record.
338	Change the primary NEC on non-rated personnel.
340	Forward to NFC all guilty courts martial findings, all NJP's which affect pay and rate, administrative actions, or restoration of above.
341	Apply data from NFC regarding above.
344	Process SSN changes.
345	Record the number of dependents residing overseas with member attached to an overseas station or ship home ported overseas.
352	Process changes to population group.
355	Process proficiency pay additions or changes.
356	Process changes which reflect the reason for unavailability of members for certain types of duty.
359	Process active duty service date changes.
362	Change a date on which member was received to a command.
376	Modify a accounting category code.
378	Process executed reserve contract extensions.
379	Align active duty obligation data with the member's service record.
382	Process a USNR agreement to remain on active duty or a USN agreement to extend an enlistment.
383	Process an operative extension of a USN member's enlistment.
385	Process the number of months of involuntary extension of a member's active duty obligation.
386	Process a cancellation of a previously executed extension of enlistment for USN or USNR members.
387	Make operative a previously executed agreement to extend the active duty of USN members.
388	Cancel a previously executed agreement to extent active duty for USNR members.
389	Make operative a previously executed agreement to extend a USNR member's enlistment.
390	Process a correction to a member's pay entry base date.
6XX	Apply activity losses to the on board or past activity data.
7XX	Same as above.
798	Cancel an activity gain reported in error.
8XX	Process strength losses from the Naval Service.
8XX	Same as above.
951	Process data which identifies those members who are considered deserters.
996	Align loss data in NFC and NMPC files.
998	Remove members who have been gained to active Naval strength in error.

APPENDIX D. INDIVIDUAL TRANSACTION ERROR RATES

1989	MONTH	JAN	FEB	MAR	APR	MAY	JUN
TAC	A68						
INPUT		477	3547	4958	27466	5543	
ERROR		32	298	102	311	528	
N1652		126	655	211	160	107	
RESEARCH		29	298	60	241	528	
TAC	QCO						
INPUT		9065	9856	9773	8871	7320	
ERROR		280	1159	733	360	318	
N47		1521	2035	1325	1294	1838	
RESEARCH		173	639	476	243	129	
TAC	300						
INPUT		4651	4282	4689	4403	5172	
ERROR		226	222	308	236	239	
N1652		135	208	408	249	217	
RESEARCH		78	91	167	135	124	
TAC	301						
INPUT		1320	827	867	1349	914	
ERROR		56	21	149	45	24	
N1652		12	18	49	57	17	
RESEARCH		55	21	19	45	24	
TAC	328						
INPUT		35576	34884	32564	30866	36769	
ERROR		2467	6193	5795	3482	3212	
N1652		731	6529	5355	1364	891	
RESEARCH		2459	6182	4593	3467	3208	
TAC	340						
INPUT		4331	2987	4162	4286	2864	
ERROR		325	247	532	482	229	
N1652		22	67	77	61	54	
RESEARCH		325	247	532	482	229	

JUL	AUG	SEP	OCT	NOV	DEC	TOTAL	ERROR%
10281	9887	38602	8367			109128	3.10%
64	1437	193	423			3388	
192	169	155	2692			4467	
64	1429	191	420			3260	
14182	11830	9252	9237			89386	12.41%
4721	2522	496	503			11092	
2536	2431	1802	1662			16444	
4190	2209	279	226			8564	
6509	5507	5917	6127			47257	4.64%
263	229	255	216			2194	
280	254	321	305			2377	
115	100	114	94			1018	
1189	1048	1404	1380			10298	5.15%
32	40	75	88			530	
22	37	37	55			304	
32	40	75	88			399	
34899	33310	37651	52508			329027	13.61%
3842	2935	3916	12950			44792	
2211	1314	1231	1373			20999	
3828	2918	3884	12900			43439	
4017	3401	3366	3545			32959	9.41%
386	264	268	368			3101	
62	71	75	78			567	
386	264	268	368			3101	

1987	MONTH	JAN	FEB	MAR	APR	MAY	JUN
TAC	A68						
INPUT			4096	189	34772		49113
ERROR			21	8	73		154
N1652			168	61	84		48
RESEARCH			19	8	73		131
TAC	QCO						
INPUT			10615	10058	10240		8098
ERROR			501	1582	656		336
N47			1236	2113	1174		1023
RESEARCH			220	309	537		171
TAC	300						
INPUT			4300	4636	7034		3882
ERROR			266	281	457		233
N1652			136	123	214		133
RESEARCH			266	281	457		233
TAC	301						
INPUT			1275	851	1232		647
ERROR			47	39	44		61
N1652			41	23	32		21
RESEARCH			41	36	44		37
TAC	328						
INPUT			28397	30650	38347		21228
ERROR			1344	1531	1960		1321
N1652			2622	765	1522		893
RESEARCH			1335	1511	1949		1316
TAC	340						
INPUT			4464	5633	6486		4446
ERROR			288	362	411		321
N1652							
RESEARCH			288	362	411		321

JUL	AUG	SEP	OCT	NOV	DEC	TOTAL	ERROR%
1434	8473	7167	27442	248	89372	222306	1.20%
141	1172	27	350	21	691	2558	
193	105	113	797	91	71	1731	
139	1171	27	311	21	669	2569	
9860	8215	11079	10580	8604	9060	96409	8.26%
303	391	3200	348	310	334	7961	
1166	998	840	1324	853	946	11673	
202	292	3113	206	239	189	5478	
4180	3965	2784	4189	5129	3112	43211	9.50%
240	260	181	331	1522	335	4106	
132	172	174	190	1063	244	2581	
240	260	181	331	1510	279	4038	
1333	952	974	1125	923	986	10298	5.33%
50	32	28	49	81	118	549	
37	25	31	27	58	39	337	
49	32	25	49	81	114	508	
83741	37083	39981	56466	29054	25570	390517	22.32%
37560	15685	11459	9778	3448	3072	87158	
9959	828	5327	3082	2238	1602	28838	
37549	15677	11428	9708	3423	3045	86941	
5053	5193	3537	5731	4918	3919	49380	7.50%
433	473	234	464	406	310	3702	
						0	
433	473	234	464	406	310	3702	

TOTAL	TOTAL		ERROR%	ERROR%
109128	222306		3.10%	1.20%
3388	2658			
4467	1731	AVG		2.15%
3260	2569			
89386	96409		12.41%	8.26%
11092	7961			
16444	11673	AVG		10.33%
8564	5478			
47257	43211		4.64%	9.50%
2194	4106			
2377	2581	AVG		7.07%
1018	4038			
10298	10298		5.15%	5.33%
530	549			
304	337	AVG		5.24%
399	508			
329027	390517		13.61%	22.32%
44792	87158			
20999	28838	AVG		17.97%
43439	86941			
32959	49380		9.41%	7.50%
3101	3702			
567	0	AVG		8.45%
3101	3702			

APPENDIX E. ERROR TRACKING SHEET

USE THIS SHEET TO RECORD THE STATUS OF ERRORS YOU RESEARCH.

1. TRANSACTION - Record the three character transaction code.
2. Check the block which refers to the transaction status.
   - CORRECT AS IS - Implies that even thought the transaction was rejected from the update, it is correct as submitted and will simply be resubmitted.
   - CORRECTED REINPUT - Implies that there was some correction which had to be made to the transaction as submitted, before it could be reinput.
   - INCORRECT NOT REINPUT - Implies the transaction was in error and will not be submitted to the update.
3. DAYS BETWEEN UPDATES - Record the number of days from the update the transaction was first submitted until the day you resubmit. (Example TC333 - TC330 = 3)

TRANSACTION	CORRECT REINPUT	CORRECTED REINPUT	INCORRECT NOT REINPUT	DAYS BETWEEN UPDATES
___				
___				
___				
___				
___				
___				
___				
___				
___				
___				
___				
___				
___				
___				

APPENDIX F. ERROR PROBABILITIES AND CORRECTION TIMES

FINDING THE ERROR PROBABILITIES

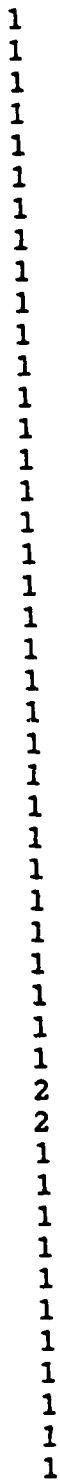
TAC	REINPUT	CORRECTD	NOTINPUT	P	P´
A68		1		1.00	0.00
QCO	14	19	81	0.88	0.12
300		3	1	1.00	0.00
301	2	4	2	0.75	0.25
328	3	1	2	0.50	0.50
340	1	16	7	0.96	0.04

DAYS UNTIL REINPUT

TAC	A68	QCO	300	301	328	340
DAYS		2		9	2	2
		1		9	2	
		2			1	
		3				
		2				
		1				
		3				
		2				
		2				
		2				
		2				
		2				
		2				
		1				
	NO DATA	1.9	NO DATA	9.0	1.7	2.0

DAYS UNTIL CORRECTED/DELETED

TAC	A68	QCO	300	301	328	340
	2	1	2	2	28	2
		2	15	1	28	2
		1	27	2		2
		1	27	1		2
		1		9		2
		1				2
		2				2
		3				2
		1				1
		2				25
		2				25
		1				25
		1				25
		1				14
		1				15
		2				15
		1				15
		1				15
		1				15
		1				
		2				
		1				
		2				
		2				
		1				
		1				
		2				
		1				
		1				
		1				
		1				
		1				
		1				
		1				
		1				
		1				
		1				
		1				
		1				
		1				
		2				
		1				
		1				
		1				
		1				
		2				



2.0	1.2	17.8	3.0	28.0	10.8

APPENDIX G. INTRINSIC TRANSACTION/STORED MIS ERROR RATES

FINDING THE STORED MIS ERROR RATE

VARIABLE	r	P	P´	ERROR RATE e(T)	
TRANSACTION					
A68	0.02	1.00	0.00	2.15%	
QCO	0.10	0.88	0.12	0.00%	NOTE 1
300	0.07	1.00	0.00	7.07%	
301	0.05	0.75	0.25	0.00%	NOTE 1
328	0.18	0.50	0.50	0.00%	NOTE 2
340	0.08	0.96	0.04	4.68%	

NOTE 1: 0 because r is less than P´
NOTE 2: P equal P´ so no valid value can be found

VARIABLE	C1	C2	C3
TRANSACTION			
A68	1.5	2.0	NO DATA
QCO	1.5	1.2	1.9
300	1.5	17.8	NO DATA
301	1.5	3.0	9.0
328	1.5	28.0	1.7
340	1.5	10.8	2.0

VARIABLE TRANSACTION	u(T)	MIS ERROR RATE e(M)	
A68	1460	0.11%	MISSING DATA
QCO	1080	0.16%	
300	1460	0.19%	MISSING DATA
301	5475	0.07%	
328	1460	0.16%	MISSING DATA
340	9000	0.02%	

LIST OF REFERENCES

Active Duty Enlisted Data Elements Catalog, Appendix A, current to change 5, 1988.

American Management Systems, Inc., for Naval Military Personnel Command, *IMPDB Logical Data Model, Revision 1* (LDM Report), 31 May 1989.

Assistant Secretary of the Navy (Financial Management), *Department of the Navy (DON) Strategic Plan for Managing Information and Related Resources (IRSTRATPLAN)* (SECNAV Instruction 5230.10), 1 April 1987.

Assistant Secretary of the Navy (Financial Management), *Information Resources (IR) Program Planning* (SECNAV Instruction 5230.9A), 16 October 1985.

Ballou, D.P., and Pazer, H.L., "Modeling and Process Quality in Multi-Input, Multi-Output Information Systems," *Management Science*, v.31, n.2, pp.150-162, February 1985.

Ballou, D.P., and Kumar.Tayi, G., "Methodology for Allocating Resources for Data Quality Enhancement," *Communications of the ACM*, v.32, n.3, pp.320-329, March 1989.

Benoit, P. S., "Handling Rejected Input Transactions", *Journal of Systems Management*, pp.26-28, May 1979.

Brancheau, J.C., and Wetherbe, J.C., "Key Issues in Information Systems Management," *MIS Quarterly*, pp.23-35, March 1987.

Brodie, M.L., "Data Quality in Information Systems," *Information and Management*, v.3, pp.245-258, 1980.

Cash, J.I., and others, *Corporate Information Systems Management, Text and Cases*, 2nd ed., Richard C. Irwin, Inc., 1988.

Commander, Naval Military Personnel Command, *Format and Procedures for Validation of Enlisted Distribution and Verification Report (EDVR)* (NAVMILPERSCOM Instruction 1080.1D), 27 January 1989.

Commander, Naval Military Personnel Command Letter (Draft) Subject: Timeliness Performance Reports, no date.

Curran, M., Enlisted Personnel Management Center, Memorandum on DMRS Timeliness, no date.

Date, C.J., *An Introduction to Database Systems*, v.1, 4th ed., Addison-Wesley Publishing Company, pp.437-460, 1986.

Davis, G.B. and Olson, M.H., *Management Information Systems: Conceptual Foundations, Structure, and Development*, 2nd ed., McGraw Hill, Inc., pp.131-160, 199-234, 501-528, 603-628, 1985.

Director, Total Force Information Resources and Systems Management Division, Chief of Naval Operations, *The MPT Information Resources Management Strategy, Volume I: Executive Overview* (MPT IRM, Volume I), February 1988.

Director, Total Force Information Resources and Systems Management Division, Chief of Naval Operations, *The MPT Information Resources Management Strategy, Volume II: The MPT Data Strategy* (MPT IRM, Volume II), 26 June 1989.

Director, Total Force Information Systems Management Department, Chief of Naval Personnel, *Component Information Resources Management Plan* (CIRMP), July 1989.

Director, Total Force Information Systems Management Department Letter 5400 Serial 16B4/0429, Subject Total Force Information Systems Management Department Reorganization (Director, Reorganization), 11 April 1988.

Deputy Chief of Naval Operations (MPT) Letter 5230 Serial 161/0591, Subject: Memorandum of Understanding (MOU) on Military Personnel and Pay Information Interface, 16 May 1988.

Deputy Chief of Naval Operations (MPT) Letter 5230 Serial 161/0593, Subject: Memorandum of Understanding (MOU) on Military Personnel and Pay Information Interface, 23 May 1989.

Deputy Chief of Naval Operations (MPT), *Manpower, Personnel and Training (MPT) Information Resources Management (IRM) Program* (OPNAV Instruction 5230.22), 6 October 1986.

Deputy Chief of Naval Operations (MPT), *MPT Information Benefits Analysis (IBA) Guideline* (DCNO, IBA Guideline), OPNAV P161-G2-89, 5 July 1989.

Deputy Chief of Naval Operations (MPT), *MPT IRM Program Data Element Standard* (DCNO, Data Element Standard), OPNAV P161-S9-89, 5 June 1989.

Deputy Chief of Naval Operations (MPT), *MPT IRM Program Data Quality Guideline* (DCNO, Data Quality Guideline), OPNAV P161-G6-88, 3 November, 1988.

DOD Compensation Office, Navy Times Pay Chart, *Navy Times*, p.30, 2 January 1989.

Haber, S.E., Segel, F., and Solomon, H., "Statistical Auditing of Large-Scale Management Information Systems," *Naval Reserve Logistics Quarterly* v.19, n.3, pp.449-459, September 1972.

Hickman, J.R., Naval Audit Service, *Military Personnel Data Branch Efficiency Review*, May 1987.

Hill, E., Total Force Information Systems Management Department, *IMPDB Logical Data Model Report* (LDM Report), 21 October 1988.

Laudon, K.C., "Data Quality and Due Process in Large Interorganizational Record Systems," *Communications of the ACM*, v.29, n.1, pp.4-18, January 1986.

Lecture by T. Abdel-Hamid, Naval Postgraduate School, 13 November 1989.

Leong-Hong, B.W. and Plagman, B.K., *Data Dictionary Directory Systems*, John Wiley & Sons, Inc., pp.25-60, 1982.

Mahmoud, E., and Rice, G., "Database Accuracy: Results from a Survey of Database Vendors", *Information and Management*, v.15, pp.243-250, 1988.

Martin, J., *Design and Strategy for Distributed Data Processing*, Prentice-Hall, Inc., pp.16-72, 375-405, 1981.

McCall, J., P. Richards, and G. Walters, *Factors in Software Quality*, 3v., NTIS AD-A049-014, 015, 055, November 1977.

McGovern, B., Navy Finance Center, Letter Code 52122, Subject NES Update Costs, w/enclosure from the Resource Accounting System (RAS), 3 November 1989.

Milestone IV System Decision Paper (SDP) for the Navy Enlisted System (NES) (Draft), no date.

Monroe, W., Total Force Information Systems Management Department, Slide Show, no page numbers, no date.

Morey, R.C., "Estimating and Improving the Quality of Information in a MIS," *Communications of the ACM*, v.25, n.5, pp.337-342, May 1982.

Navy Finance Center, "Creation and Reconciliation," *JUMPS Design Requirement Manual* (JDRM), v.II, 1990.

Pressman, R.S., *Software Engineering A Practitioner's Approach*, 2nd ed., McGraw-Hill Inc., pp.433-463, 526-550, 1987.

Software Solutions, Inc., *Software Architecture Level I Document* (Draft), 31 May 1988.

Teter, C., Navy Finance Center (Code 6C), Pay/Personnel Interface Meeting Issue Subject: Timeliness Measurement Review, 26-28 July 1989.

Tidewater Consultants, Inc., for Naval Military Personnel Command, *Revised Integrated Military Personnel Data Base (IMPDB) General Functional Requirements*, August 1989.

Troy Systems, Inc., for Naval Military Personnel Command, *Data Quality Improvement Report*, 29 September 1989.

Troy Systems, Inc., for Naval Military Personnel Command, *Error Research and Corrections Analysis*, 7 July 1989.

U.S. Department of Commerce, NBS Special Publication 500-13, *Features of Seven Audit Software Packages--Principles and Capabilities*, by A.J. Neumann, July 1977.

U.S. Office of Personnel Management, *Position Classification Standard for Auditing Series GS-511*, May 1982.

U.S. Office of Personnel Management, *Position Classification Standard for Computer Specialist Series GS-334*, December 1980.

Valett, J.D., and McGarry, F.E., "A Summary of Software Measurement Experience in the Software Engineering Laboratory," *The Journal of Systems and Software*, v.9, pp.137-148, 1989.

Varley, Thomas C., *Data Input Error Detection and Correction Procedures*, Ph.D. Dissertation, George Washington University, Office of Naval Research Report, Serial T-222, 2 June 1969.

Weber, R., *EDP Auditing Conceptual Foundations and Practice*, Mc Graw-Hill, Inc., pp.3-20, 164-182, 207-248, 1982.

Weekly Federal Employees' News Digest, January 1989 General Schedule Pay Chart, copy of page with no date.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Technical Information Center Cameron Station Alexandria, VA 22304-6145	2
2.	Library, Code 0142 Naval Postgraduate School Monterey, CA 93943-5002	2
3.	Henry Blankinship (NMPC-1651) Navy Department, Federal Building 2 Naval Military Personnel Command Washington, DC 20370-5000	1
4.	Mary Curran Enlisted Personnel Management Center New Orleans, LA 70159-7900	1
5.	PNCM(SW) J. Gasch (NMPC-453) Navy Department, Federal Building 2 Naval Military Personnel Command Washington, DC 20370-5000	1
6.	Peggy Gross (NMPC-1641E) Navy Department, Federal Building 2 Naval Military Personnel Command Washington, DC 20370-5000	1
7.	LCDR P. Little (OP-164) Navy Department, Federal Building 2 Naval Military Personnel Command Washington, DC 20370-5000	1
8.	Willie Monroe (NMPC-1653F) Navy Department, Federal Building 2 Naval Military Personnel Command Washington, DC 20370-5000	1
9.	PNC R. Morrow (NMPC-1641E) Navy Department, Federal Building 2 Naval Military Personnel Command Washington, DC 20370-5000	1
10.	LT Susan R. Sablan 2540A South Walter Reed Drive Arlington, VA 22206	2